



PHD

Numerical analysis of dynamical systems

Humphries, Antony R.

Award date:
1993

Awarding institution:
University of Bath

[Link to publication](#)

Alternative formats

If you require this document in an alternative format, please contact:
openaccess@bath.ac.uk

Copyright of this thesis rests with the author. Access is subject to the above licence, if given. If no licence is specified above, original content in this thesis is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC-ND 4.0) Licence (<https://creativecommons.org/licenses/by-nc-nd/4.0/>). Any third-party copyright material present remains the property of its respective owner(s) and is licensed under its existing terms.

Take down policy

If you consider content within Bath's Research Portal to be in breach of UK law, please contact: openaccess@bath.ac.uk with the details. Your claim will be investigated and, where appropriate, the item will be removed from public view as soon as possible.

Numerical Analysis of Dynamical Systems

Submitted by

Antony R. Humphries

for the degree of PhD

of the

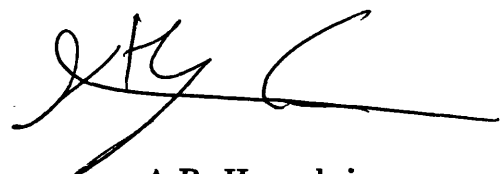
University of Bath

1993

COPYRIGHT

Attention is drawn to the fact that copyright of this thesis rests with its author. This copy of the thesis has been supplied on the condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without the prior written consent of the author.

This thesis may be made available for consultation within
the University Library and may be photocopied or lent to other libraries
for the purposes of consultation.

A handwritten signature in black ink, appearing to be 'A.R. Humphries', with a long horizontal stroke extending to the right.

A.R. Humphries

UMI Number: U540493

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI U540493

Published by ProQuest LLC 2013. Copyright in the Dissertation held by the Author.
Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against
unauthorized copying under Title 17, United States Code.



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

UNIVERSITY OF BATH LIBRARY	
22	14 JUL 1994
P40	

5083668

Dedicated to
Mary L. Yorke
and
any other teacher who has
ever inspired a pupil.

Summary

This thesis is concerned with the numerical solution by fixed time-stepping methods of finite dimensional dynamical systems over arbitrarily long time intervals. We treat the step-size h as a continuation parameter and study what happens as h is refined. There are two introductory chapters. In Chapter 2 dynamical systems theory is reviewed and classes of dissipative and gradient systems are introduced. In Chapter 3 Runge-Kutta and linear multistep methods are defined and many preliminary results are proved; including solubility results for implicit methods under the structural conditions that we consider. Chapter 4 contains a study of *spurious solutions*. The existence and boundedness of spurious fixed points and two-cycles for Runge-Kutta and linear multistep methods is studied in the limit as $h \rightarrow 0$. Amongst other results it is shown that if f is Lipschitz then spurious solutions become unbounded as $h \rightarrow 0$ (if they exist for h arbitrarily small). In Chapter 5 the numerical solution of *gradient systems* by Runge-Kutta methods is studied. We consider under what conditions the numerical solution defines a discrete gradient system. In Chapter 6 the numerical solution of two classes of *dissipative* system is considered. Conditions are derived under which the numerical solution preserves the dissipativity of the underlying system; algebraically stable methods in particular are seen to perform well. In the final chapter the *set convergence* of numerical approximations, using Runge-Kutta methods, to attractors and invariant sets of dynamical systems as $h \rightarrow 0$ is considered. A new method of proof of *lower semicontinuity* is introduced which enables Hausdorff convergence to be proved for some non-gradient systems.

Acknowledgements

I have many people to thank:

First and foremost Andrew Stuart for all his help, encouragement, advice, and even putting the boot in occasionally when it was needed. Certainly the world's best supervisor.

Arieh Iserles for introducing me to this subject, the various (and too numerous to name) people with whom I have had useful discussions, and Andy Wathen and Chris Budd for being so patient when I overran.

Everyone at Bath who made my time there what it was; especially Ching Lee, Rob Douglas, Bath Bridge Club, Minerva-Bath Rowing Club and Carole Sheringham-Smith.

At Stanford, Mary Washburn and family for being so hospitable and even "adopting" me. Also Margot Gerritsen, Gene Golub, Kjell and Eva Gustafsson, Stanford Fencing and the rest of the SCCM group.

I must not forget the mob from Cambridge (especially Aino Winther-Pedersen and Liz Harvey; now Liz Marley, I have taken so long to write up).

Finally and perhaps most importantly my family, without whom I would not be here. (What time is dinner mum ?)

I received financial support from the Science and Engineering Research Council with additional support for my visit to Stanford from Stanford University itself and the US Office of Naval Research (grant N0014-92-J-1876).

Tony Humphries

Bristol Dec '93

Contents

1	Introduction	7
1.1	Dynamical Systems	7
1.2	The Problem	10
1.3	Outline and Main Results	12
1.4	Future Research and Other Related Work	23
2	Dynamical Systems	25
2.1	Introduction and Lipschitz Continuity	25
2.2	Dissipative Dynamical Systems	33
2.2.1	Structural Assumptions Inducing Dissipativity	33
2.2.2	Absence of One-sided Lipschitz Condition	37
2.3	Invariant Sets and Attractors	42
2.4	Gradient Dynamical Systems	49
2.5	Discrete Dynamical Systems	53
2.5.1	Discrete Gradient Systems	54
2.5.2	Implicit Maps	55
2.6	Fixed Point Theorems	56
3	Numerical Methods	58
3.1	Introduction	58
3.2	Runge-Kutta Methods	59
3.2.1	Reducibility	61
3.2.2	Properties of Runge-Kutta Solutions	62
3.3	Linear Multistep and One-Leg Methods	65
3.4	Numerical Stability Theories	67
3.5	Examples	69

3.6	Runge-Kutta Methods as Dynamical Systems	72
3.6.1	Solubility of Implicit Runge-Kutta Equations under Continuity Conditions	72
3.6.2	Local and Global Error Bounds	76
3.6.3	Solubility of Implicit Runge-Kutta Equations under Structural Assumptions	82
3.7	Linear Multistep Methods as Dynamical Systems	94
3.7.1	Solubility of Linear Multistep Defining Equations	95
4	Spurious Limit Sets	96
4.1	Introduction	96
4.2	Spurious Fixed Points of Runge-Kutta Methods	101
4.3	Spurious Two-Cycles of Runge-Kutta methods	109
4.4	Spurious Two-Cycles of Linear Multistep Methods	114
4.5	Spurious Solutions of Nonautonomous Systems	118
5	Gradient Systems	123
5.1	Introduction	123
5.2	Approximation of Lipschitz Gradient Systems	125
5.3	One-sided Lipschitz Gradient Systems	130
6	Dissipative Systems	135
6.1	Introduction	135
6.2	Dissipativity of Algebraically Stable Methods	139
6.3	Dissipativity Under Global Lipschitz Condition	144
6.4	Dissipativity of Theta Methods	149
7	Attractors and Invariant Sets	152
7.1	Introduction	152
7.2	Upper Semicontinuity	157
7.3	Lower Semicontinuity	160
7.4	Invariant Sets and Attractors	166
	Bibliography	177

Chapter 1

Introduction

1.1 Dynamical Systems

Many interesting problems from such diverse fields as physics, engineering, biology and economics are modelled by initial value problems which give rise to systems of ordinary differential equations or *dynamical systems*. A complete theory exists for the solution of linear ordinary differential equations, but many realistic models are nonlinear. These nonlinear systems are, in general, nonintegrable; that is there are no analytical closed form solutions. Perturbation and averaging techniques were developed to treat weakly nonlinear problems, and Poincaré developed tools for the qualitative solution of strongly nonlinear systems, but with notable exceptions the quantitative solution of strongly nonlinear systems had to await the advent of the computer age.

In a computer simulation a numerical method is used to discretize the differential equation, replacing it with a finite dimensional map, and this map is then iterated on the computer. The essential question to consider then is

What is the relationship between the flow generated by the underlying differential equation and the flow generated by the map used to model the system numerically ?

Classical convergence results for numerical methods provide the answer to this question over finite time intervals. These results give error bounds of the form

$$e^{cT} h^p$$

for individual trajectories, where h is the step-size and $p \geq 1$ is the order of the method,

c is a (typically positive) constant and T is the length of the time interval over which the integration occurs. Such estimates can be used to show that, on a compact time interval, the trajectory generated by the map converges to the corresponding trajectory of the underlying system as $h \rightarrow 0$.

Although transient behaviour can be important, in dynamical systems theory it is often the long term or asymptotic behaviour of the system that is of most interest. For example, a biologist may not be very interested in minor fluctuations in the elephant population, but would be very interested to know that the population will explode or that the elephant will become extinct. Similarly, mathematicians and physicists have long pondered the stability of the solar system; “can the earth crash into the sun ?” is not as silly a question as it might at first sound.

As $T \rightarrow \infty$ classical error estimates become unbounded (except in rare cases where $c < 0$; for example when the solution converges to a hyperbolic equilibria, in which case uniform in time error estimates can be derived; see Stetter [50] and Sanz-Serna and Stuart [49]) and methods for initial value problems which are convergent in finite time do not necessarily yield the same asymptotic behaviour as the underlying differential equation for small fixed step-size.

The asymptotic behaviour of many systems is confined to a bounded set. Such systems are, in dynamical systems parlance, said to be *dissipative*, and the compact set which contains all the asymptotic behaviour of the system is called the *global attractor*. The most interesting attractors are the so-called *strange* or *chaotic* attractors. On such an attractor trajectories typically diverge exponentially in time, and if a trajectory on a strange attractor is approximated numerically then small numerical errors will be amplified at each step. Under such conditions the error itself, not just the error bound, will become large.

An approximation to the attractor is often obtained by integrating the system numerically over a very long time interval and plotting the resulting trajectory, after discarding the initial transient phase. Since the error bound for the numerical solution becomes unbounded as $T \rightarrow \infty$ there is no reason to suppose that this procedure produces a good approximation to the attractor. However in practice different numerical methods with different step-sizes, provided a sufficiently small step-size is chosen, seem to result in remarkably similar “pictures” of the attractor. This seems to suggest that the numerical solutions are indeed providing good approximations to the attractors of

the underlying systems. However theory has lagged behind simulation, and until recently there has been little rigorous theory to show that these numerical solutions are good approximations to the nonlinear system which they are supposed to model. Dynamicists, to their credit, usually note this problem, and even have a heuristic argument to show why the approximation should be good. Ian Stewart [51] gives the following argument for numerical solution of the Lorenz equations, but it applies equally well to other systems:

If you think you're solving the initial value problem for the Lorenz equations, with the exact numerical conditions that you fed into the computer, then you're fooling yourself. But if you think you're plotting out the shape of the attractor rather than a trajectory on it, you're in good shape. Tiny errors that move your point away from the attractor rapidly die out - that's what attractor means. It's only errors that stay on the attractor that blow up. that's the argument; it seems to work. But it's by no means watertight.

Perhaps not; but it can be made rigorous. In Theorem 7.2.2 we will show that a numerical approximation to a dynamical system with an attractor \mathcal{A} , itself possesses a (numerical) attractor \mathcal{A}_h , if the step-size h of the method is sufficiently small, and moreover that \mathcal{A}_h converges to \mathcal{A} in a set-theoretic sense as $h \rightarrow 0$. The proof of this result follows the method of Hale, Lin and Raugel [29] who proved a similar result for perturbations of infinite dimensional systems. It is interesting to note that the formal proof is very close in spirit to the heuristic argument.

We will not solve any new dynamical systems in this work. Seemingly uncountable numerical simulations of nonlinear dynamical systems have already been conducted. Rather than perform more such simulations, it is our aim to develop a framework and rigorous theory which can be used to compare the dynamics of numerical approximations with the dynamics of the underlying system over arbitrarily long time intervals. Hence we hope to alleviate the need for heuristic arguments, such as the one stated above, wherever possible, and at least to some extent, to justify the validity of these numerical simulations.

Due to the exponential divergence of trajectories on a chaotic attractor we cannot expect the numerical solution to track a trajectory of the underlying system over an infinite time interval and we need not only to develop new techniques to compare the discrete and underlying systems, but we need also to find a new way of thinking

about the problem. Under suitable conditions the numerical solution defines a discrete dynamical system in its own right and this will be the key to our approach. Instead of comparing individual trajectories, we will compare and contrast the features of the underlying dynamical system with those of the discrete dynamical system defined by its numerical approximation. This will enable us to compare the asymptotic behaviour of the two systems.

1.2 The Problem

We consider the numerical approximation, over arbitrarily long time intervals, of autonomous first order dynamical systems defined by

$$\frac{d\mathbf{y}}{dt} = \mathbf{f}(\mathbf{y}) \quad (1.2.1)$$

for $t \in [0, \infty)$, $\mathbf{f}: \mathbb{R}^m \rightarrow \mathbb{R}^m$, $\mathbf{y}(t) \in \mathbb{R}^m$ and an arbitrary initial condition $\mathbf{y}(0) = \mathbf{y}_0$. We will also sometimes consider the backwards in time solution, in which case t will be negative.

Equation (1.2.1) may arise in many ways, but we will usually think of (1.2.1) as a spatial semi-discretization of a partial differential equation, in which case m may be large. We will not consider partial differential equations directly in this thesis, but will assume that under suitable discretization (1.2.1) inherits certain structure from the partial differential equation, and we will then seek a numerical solution to (1.2.1) that preserves this structure. We will consider several different structural assumptions on \mathbf{f} in this thesis.

We will consider fixed time-stepping numerical methods throughout, and will concentrate on Runge-Kutta methods. Linear Multistep methods will also be considered in Chapter 4. Many of the ideas, and some of the results, presented here can be extended to variable time-stepping Runge-Kutta methods; see Stuart and Humphries [55].

In (1.2.1) we have limited attention to first order systems. This is not a restriction, since by introducing additional variables higher order systems can be converted to first order systems. Indeed conversion to a first order system is often the first stage in the analysis of a nonlinear system for dynamicists and numerical analysts alike.

We will often require a norm $\|\bullet\|$ on \mathbb{R}^m , and sometimes also an inner product $\langle \bullet, \bullet \rangle$. Unless otherwise stated the inner product on \mathbb{R}^m is arbitrary. Where we also

use an inner product the norm used is the one defined by the inner product, but if an inner product is not required then the norm is arbitrary unless otherwise stated.

We will state continuity and differentiability conditions for \mathbf{f} as they are required. It will often be assumed that \mathbf{f} is merely locally Lipschitz; and weaker continuity conditions are also considered.

There are two complimentary approaches for studying the asymptotic behaviour of (1.2.1); following Beyn [5], these are often referred to as indirect and direct methods.

In the *indirect method* a numerical integration routine is used to solve (1.2.1). If (numerical) solution trajectories are plotted for several initial conditions then a good picture of the dynamics can be built up.

In the *direct method* approach instead of solving for individual trajectories of (1.2.1), defining equations are set up and solved to identify sets which are invariant under the evolution of (1.2.1).

Together these two approaches can give a detailed picture of the dynamics of (1.2.1), *but* there are serious problems with each approach.

The indirect method is very easy to perform, but as already noted, classical error bounds typically become unbounded as $t \rightarrow \infty$, and so there is no a priori reason why the numerical trajectories should approximate trajectories of the underlying system. The picture of the dynamics that the indirect method gives is for the discrete dynamical system defined by applying the numerical method to the underlying dynamical system, *not* for the underlying dynamical system itself. To justify results obtained by the indirect method we need to examine the relationship between the underlying dynamical system and the discrete dynamical system defined by its numerical approximation.

Direct methods can directly identify the invariant sets of (1.2.1), and, in general, bounds can be derived for the distance between the set identified by the numerical solution of the defining equation and the set defined by its exact solution. Nevertheless there is a fundamental problem with this approach; only invariant sets for which defining equations can be written down can be identified. Whilst $\mathbf{f}(\mathbf{x}) = 0$ is the defining equation for the fixed points of (1.2.1), and defining equations can also be written down for periodic orbits, homoclinic orbits and tori, it seems inconceivable that defining equations can be written down for more complicated invariant sets such as strange attractors.

Thus the problem with the direct approach is that complicated invariant sets, and

chaotic dynamics might go unobserved. There is a danger of using the defining equations at hand and then only seeing what is looked for, and overlooking invariant sets for which there are no (known) defining equations. For this reason we prefer the indirect approach which can be used to give a global picture of the dynamics, and will consider this approach throughout.

Under fairly general conditions, which we will set out later, the numerical approximation using a fixed time-stepping method defines a discrete dynamical system. Indeed the discretization defines a whole family of discrete dynamical systems, parameterized by the step-size h used for the numerical solution. As already noted, to justify the validity of results obtained using the indirect method we must compare the dynamics of the underlying system with the dynamics of these discrete dynamical system, and this is the task that we set out to undertake.

1.3 Outline and Main Results

The main results of this thesis are contained in the last four chapters. In Chapter 4 we study *spurious solutions* of numerical methods for (1.2.1). In Chapters 5 and 6 we compare the dynamics of numerical solutions with the dynamics of (1.2.1) for two classes of dynamical system; namely *gradient* and *dissipative* systems. Finally in Chapter 7 we consider the set convergence of numerical attractors to attractors of (1.2.1).

Before we can set out our main results we need two preliminary chapters on dynamical systems and numerical methods in which we introduce concepts from these two fields and prove many preliminary results, which are essential for the proofs of our main results in later chapters. Known results from the literature, which are required in the exposition, will be labelled “Result” to distinguish them from the original results herein which are variously labelled as lemmas, propositions, theorems and corollaries. Proofs will be given for some “Result”s, where these are not readily available in the literature. We will mention relevant articles in this section, however each chapter contains its own introduction and more citations and/or more detail on the citations below can be found in those introductions.

In Chapter 2 we introduce the basic concepts such as the evolution operator, which we require from dynamical systems theory. There are many good books on this subject, including those of Hirsch and Smale [34], Guckenheimer and Holmes [24] and Wiggins [59] on finite dimensional systems and the books of Hale [28] and Temam [58]

on infinite dimensional systems. We also define the two classes of dynamical system which we consider in Chapters 4 and 5; *gradient systems* have the property that all trajectories of the system are asymptotic to fixed points, whilst for *dissipative systems* the asymptotic behaviour is confined to a bounded set, but no restriction is placed on the dynamics within this bounded set. We will also introduce structural assumptions on f which imply dissipativity and gradient structure.

In Chapter 3 we introduce and study the numerical methods which will be used for the numerical solution of (1.2.1); namely fixed time-stepping Runge-Kutta, one-leg and linear multistep methods. Detailed accounts of these methods, with emphasis on their behaviour over finite time intervals, can be found in the books of Butcher [9], Dekker and Verwer [13], Hairer, Norsett and Wanner [26] and Hairer and Wanner [27]. To compare the discrete dynamical system defined by the numerical approximation with the underlying system we will often take a geometric approach, and will find that inequalities and identities satisfied by the numerical solution will often be more important than classical error bounds. In Section 3.2.2 we establish some inequalities and identities for Runge-Kutta methods, which do not in themselves appear startling, but which nevertheless represent the vital first step in the proofs of many of the results in Chapters 4, 5 and 6.

In Section 3.4 we briefly review the classical stability theories of A-stability, introduced by Dahlquist [10], for linear problems, and G-, B- and algebraic-stability for nonlinear systems.

The test problem which defines A-stability is linear, and hence its dynamics are essentially trivial and it is not very interesting from a purely dynamical systems context. The contractivity condition which is used in the G- and B-stability theories is studied in Section 2.1. There it is shown that the fixed points of such a system define a convex set. However, generically fixed points of dynamical systems are isolated, so from a dynamical systems viewpoint this is also an unnatural class of systems to consider.

Although the systems used in the definitions of the classical nonlinear stability concepts have essentially trivial dynamics, it must be emphasised that the stability concepts themselves are far from trivial. Not only do these nonlinear stability theories represent the first systematic study of numerical solutions over arbitrary long time intervals, they have been seen to have implications for the numerical solution of dynamical systems for which the dynamics are far from trivial. Our study of the dy-

namics of the numerical solutions to dissipative and gradient systems could be thought of as extending these classical nonlinear stability concepts to systems with non-trivial dynamics and a wider range of applications; this is the approach taken in Stuart and Humphries [56] where the relationship between the classical stability theories and stability theories for dynamical systems with more complicated dynamics is explored.

In Section 3.6 we consider the solubility of the equations which define the Runge-Kutta method at each step. This is necessary for implicit methods, since if there is not a unique solution of the Runge-Kutta defining equations at each step then the numerical method does not define a discrete dynamical system when applied to (1.2.1); and it then becomes meaningless to talk about comparing the discrete dynamical system defined by the numerical solution with the underlying system.

In Section 3.6.1 we consider the solubility of the Runge-Kutta defining equations under continuity conditions. This problem was first considered by Butcher [7], under the assumption that \mathbf{f} is globally Lipschitz. We give simple extensions of Butcher's result to weaker continuity conditions and in Theorem 3.6.4 also show that if there is an a priori bound on the numerical solution which implies that some bounded set is forward invariant under the evolution of the numerical approximation, then the numerical solution defines a discrete dynamical system on this set for h sufficiently small. The structural assumptions which we impose on \mathbf{f} sometimes imply such a bound.

In Section 3.6.3 we consider the solubility of the Runge-Kutta defining equations under the various structural assumptions that we consider. This problem has been studied extensively in the stiff numerical analysis literature under the assumption that \mathbf{f} satisfies a one-sided Lipschitz condition, see [13, 27]. We will use the notation and ideas from this theory to study the existence of solutions to the Runge-Kutta defining equations under the dissipativity inducing structural assumptions that we consider. We prove existence of solutions to the Runge-Kutta defining equations for any step-size $h > 0$,

- for a large class of Runge-Kutta methods, including many algebraically stable methods, when \mathbf{f} satisfies (2.2.1) (see Theorem 3.6.18),
- and for a smaller class of Runge-Kutta methods when \mathbf{f} satisfies (2.2.9) (see Theorem 3.6.20).

However, when \mathbf{f} does not satisfy a one-sided Lipschitz condition, the solution of the

Runge-Kutta defining equations will not in general be unique; in Example 3.6.19 we show how to construct multiple solutions for the backward Euler method, when f satisfies either of the structural conditions considered.

Whilst the nonlinear stability theories mentioned above, could be considered to be the first systematic study of the numerical analysis of (a class of) nonlinear dynamical systems, this area has been attracting increasing attention since the mid 1980's and is now a very busy field of research. The review articles of Beyn [5], Sanz-Serna [48] and Stuart [54] give a good indication of most of the directions being pursued. We will highlight some of these below.

In Chapter 4 we begin our study of the dynamics of numerical solutions to (1.2.1) in earnest by studying the existence and behaviour of spurious solutions. Iserles [38] showed that Runge-Kutta and linear multistep methods retain all the fixed points of (1.2.1), however some Runge-Kutta methods (but not linear multistep methods) may generate additional fixed points which do not correspond to fixed points of (1.2.1). These additional steady solutions introduced by the discretization are referred to as *spurious fixed points*. Some Runge-Kutta and linear multistep methods also admit solutions of the form $y_{2n} = u$, $y_{2n+1} = v \forall n \geq 0$, where $u \neq v$; known as a *period two solutions*, such periodic motion on the grid scale must also be spurious. If the numerical approximation admits spurious fixed points or period two solutions then the asymptotic behaviour of the numerical solution will differ from the asymptotic behaviour of (1.2.1) for at least some initial conditions. If the spurious solutions are stable then they may attract a large set of initial conditions, and in such a case the numerical approximation ceases to be an “approximation” to the underlying system over long time intervals. Although unstable spurious solutions will not attract a large set of initial data, it has been observed [17, 53] that the unstable manifold of the spurious solution is often connected to infinity, and thus the existence of an unstable spurious solution will cause the numerical solution to blow up for some initial conditions, and will thus destroy the structure of the underlying system. Examples of spurious fixed point and period two solutions, and their effect on the dynamics of the numerics can be found in [23, 38, 39, 46, 53, 57].

In [25, 38, 39, 40, 57] a thorough study of Runge-Kutta and linear multistep methods is conducted, in order to classify the methods which do/do not admit spurious fixed points and period two solutions. However, although many popular numerical methods

are seen to admit spurious solutions in theory, in practice these methods often perform well, and it is thus important to study the spurious solutions of these methods, rather than to simply classify the methods which do not admit spurious solutions. For this reason we study the spurious solutions of Runge-Kutta and linear multistep methods (rather than the methods themselves), and using the step-size h as a bifurcation, or continuation, parameter we study the behaviour of the spurious solutions in the limit as $h \rightarrow 0$ when simple continuity conditions are applied to \mathbf{f} . A first result along these lines is contained in Stuart and Peplow [57] who show that if $\mathbf{f} \in \mathcal{C}^1(\mathbb{R}^m, \mathbb{R}^m)$ then period two solutions of the two-stage theta method (3.5.2) become unbounded as $h \rightarrow 0$. This result is the inspiration of the work in Chapter 4. In summary, our main results are that for Runge-Kutta methods and a large class of linear multistep methods, spurious fixed points and period two solutions

- do not exist for $h \in (0, h_0)$ for some $h_0 > 0$ if \mathbf{f} is globally Lipschitz,
- become unbounded as $h \rightarrow 0$, if they exist for h arbitrarily small, if \mathbf{f} is locally Lipschitz,
- either become unbounded or converge to a true fixed point of (1.2.1) as $h \rightarrow 0$ if \mathbf{f} is continuous, but not Lipschitz continuous.

These results suggest that spurious solutions will not degrade the numerical solution on a bounded set when \mathbf{f} is Lipschitz and the step-size is sufficiently small, but that the discrete system defined by the numerical approximation on the whole of \mathbb{R}^m can possess spurious fixed points and period two solutions for h arbitrarily small. Thus we can derive good local numerical approximations to (1.2.1), but not a good global approximation to the dynamics of (1.2.1). However, in Chapter 4, we only apply continuity conditions to \mathbf{f} , and do not impose any structure on (1.2.1). We will see later that we can obtain a good global approximation to the dynamics of (1.2.1) when structure is imposed on the system.

In addition to the main results, we give sufficient step-size bounds to prevent spurious solutions when \mathbf{f} is globally Lipschitz, and to exclude spurious solutions from a bounded set when \mathbf{f} is locally Lipschitz, and also give a necessary condition for spurious solutions to bifurcate from a fixed point of (1.2.1) at $h = 0$. In Example 4.3.5 a continuous initial value problem that generates bounded spurious solutions for h arbitrarily small is presented, showing that the last of our main results is relevant, and in

Example 4.4.2 it is shown that spurious solutions which are independent of the step-size can be constructed for linear multistep methods not in the class covered by our results; so that these results cannot be extended to all linear multistep methods.

In Section 4.5 we will consider spurious solutions of nonautonomous systems. We will show that methods which do not admit spurious solutions for autonomous systems can admit spurious solutions when applied to nonautonomous problems, and we illustrate the issues involved in extending our results to this case.

Having compared the fixed points of numerical approximations with those of the underlying system (1.2.1) in Chapter 4, a natural next step is to compare the dynamics of numerical approximations with the dynamics of (1.2.1), for systems whose trajectories are all asymptotic to fixed points. Gradient systems have this property, and in Chapter 5 we consider the numerical solution of a general class of gradient systems when \mathbf{f} is either locally or globally Lipschitz or satisfies a one-sided Lipschitz condition.

The dynamics of gradient systems are studied in [28, 34]. These systems are interesting for several reasons. Firstly the Cahn-Hilliard equation [16], which models the process of coarsening in solid phase separation, and scalar reaction-diffusion equations [17] give well-studied examples of partial differential equations which are in gradient form. Under suitable spatial discretization these systems generate gradient systems of a similar form to those studied here [16, 17]. A second more philosophical reason for studying the dynamics of numerical solutions of gradient systems is because these systems do *not* display chaos. If the numerical solution of a dynamical system (1.2.1) displays chaotic dynamics then it is obviously important to know whether this is a feature of the underlying system or whether the chaos is numerically generated. If numerical solutions of gradient systems displayed chaotic dynamics then this would cast severe doubt on the validity of any numerically observed chaotic dynamics.

The main results in Chapter 5 for the numerical solution of a general class of gradient systems by Runge-Kutta methods are that

- if \mathbf{f} is globally Lipschitz then there exists $h_0 > 0$ such that if $h \in (0, h_0)$ then the numerical solution defines a discrete gradient system on \mathbb{R}^m with the same Lyapunov functional and fixed points as the underlying system.
- if \mathbf{f} is locally Lipschitz then given any bounded set B there exists $h_0 > 0$ such that if $h \in (0, h_0)$ then the numerical solution defines a discrete gradient system on a set containing B with the same Lyapunov functional and fixed points as the

underlying system.

- if f satisfies a one-sided Lipschitz condition (2.1.6) then for $h \in (0, 1/c)$ and $\theta \in [1/2, 1]$ the one- and two-stage theta methods (3.5.1) and (3.5.2) define discrete gradient systems on \mathbb{R}^m which possess the same fixed points as the underlying system and have Lyapunov functionals F_h which are an $O(h)$ perturbation of F .

These results show that when a gradient system is modelled numerically by a Runge-Kutta method, the numerical solution itself defines a discrete gradient system, under fairly general conditions. Thus under these conditions all trajectories of the numerical solution are asymptotic to fixed points of the underlying system; and there is certainly no numerically generated chaos.

The third result is related to work of Elliott and Stuart [17]. There it is shown that solution of a class of gradient systems which satisfy a one-sided Lipschitz condition by any of the first three backward differentiation formulae defines a continuous discrete gradient system with the same fixed points as the underlying system and with a Lyapunov functional which is a perturbation of the Lyapunov functional of the underlying system.

In Chapter 6 we generalize our theory by considering dynamical systems for which trajectories need not be asymptotic to fixed points, but which possess a bounded *absorbing set* which all trajectories enter in a finite time and thereafter remain inside. Recall that such systems are said to be *dissipative*. We consider the numerical solution of two classes of dissipative system defined by (2.2.1) and (2.2.9).

Many well known systems, such as the Lorenz equations are dissipative. Perhaps more importantly, dissipative systems also arise through the spatial discretization of some partial differential equations; the Cahn-Hilliard equation, the Navier-Stokes equations in two dimensions, the complex Ginzburg-Landau equation, and the Kuramoto-Sivashinsky equation all satisfy an infinite dimensional analogue of (2.2.1) [58]. Under suitable spatial discretization they generate systems of the form (1.2.1, 2.2.1).

Although the asymptotic behaviour of a dissipative system must be confined to a bounded absorbing set, we emphasise that these systems can display a variety of interesting dynamical features ranging from multiple competing equilibria (the Cahn-Hilliard equation; a gradient system) through periodic and quasi-periodic behaviour (the complex Ginzburg-Landau equation) to chaos (the Kuramoto-Sivashinsky and Lorenz equations).

We will seek to establish conditions under which the numerical solution using a Runge-Kutta method preserves the dissipativity of the underlying system, since if the absorbing set is destroyed by the discretization then incorrect asymptotic behaviour will be observed for at least some initial conditions.

The solution of “stiff” initial value problems has been widely studied in the numerical analysis literature under the assumption that f satisfies a one-sided Lipschitz condition. If we could consider dissipative systems under this assumption, then we could make use of the “stiff” theory to simplify our analysis. However we present several examples in Section 2.2.2 which show that dissipative systems in general, and the Lorenz equations in particular, do not satisfy one-sided Lipschitz conditions, and for this reason, we will not assume a one-sided Lipschitz condition when studying dissipative dynamical systems. However the lack of a one-sided Lipschitz condition means that solutions to the Runge-Kutta defining equations are not in general unique, and hence that the numerical solution does not define a discrete dynamical system. This forces us to generalize the concept of dissipativity to cover multi-valued maps.

To prove the existence of an absorbing set often requires a step-size bound which is dependent on the initial data; however absorbing sets with step-size bounds independent of the initial data have been constructed in [17, 18, 19, 43] for spatial semi-discretizations of partial differential equations satisfying infinite dimensional analogues of (2.2.1) *and* for temporal discretization of the resulting ordinary differential equations (which are of the form (1.2.1,2.2.1)). The full discretizations considered in these papers are all of first or second order in time and often correspond to applying the backward Euler method to the appropriate semi-discretized system. The main result in Chapter 6 is Theorem 6.2.2 which states that

- the numerical approximation to (1.2.1,2.2.1) defined by any algebraically stable Runge-Kutta method is dissipative (in the generalized sense of multi-valued maps) for *any* fixed step-size $h > 0$.

Since there exist algebraically stable Runge-Kutta methods of arbitrarily high order, this result shows that we can approximate (1.2.1,2.2.1) using methods of arbitrarily high order, whilst still retaining the dissipativity of the underlying system. This shows that not only the backward Euler method, but any algebraically stable Runge-Kutta method can be used to solve the spatially semi-discretized systems mentioned above, and the dissipativity of the system will be retained.

We also present an example which shows that when (1.2.1,2.2.1) is approximated by a non A-stable method then a step-size bound dependent on initial data will be required. This shows that algebraic stability is a necessary condition and A-stability is a sufficient condition for a Runge-Kutta method to preserve the dissipativity of (1.2.1,2.2.1).

Although the numerical solution does not define a discrete dynamical system on \mathbb{R}^m , Theorem 6.2.3 shows that if the step-size is sufficiently small then it does define a continuous discrete dynamical system in a natural way on its absorbing set, and that the numerical solution has the same fixed points as the underlying system (1.2.1,2.2.1).

In Section 6.3 we consider (1.2.1,2.2.1) under the additional assumption that f is globally Lipschitz. Theorem 6.3.1 shows that for these systems the numerical solution by any Runge-Kutta method with positive weights defines a continuous dissipative discrete dynamical system if h is sufficiently small. We also present an example of a globally Lipschitz dissipative system not of the form (1.2.1,2.2.1) for which the numerical solution by the forward Euler method is not dissipative for any $h > 0$. This implies that Theorem 6.3.1 cannot be extended to arbitrary globally Lipschitz dissipative systems.

In Section 6.4 we consider the numerical approximation of dissipative systems of the form (1.2.1,2.2.9) using Runge-Kutta methods. Although the results in Section 3.6.3 imply that the Runge-Kutta defining equations are soluble for the methods that we consider in this section, Example 3.6.19 again implies that this solution need not be unique, and so we must use the generalized concept of dissipativity for multi-valued maps. We show that

- the numerical solution of (1.2.1,2.2.9) defined by either the one- or two-stage theta method (3.5.1) or (3.5.2) with $\theta \in [1/2, 1]$ is dissipative in this generalized sense for any step-size $h > 0$.

Since the two-stage theta method (3.5.2) is A-stable but not algebraically stable for $\theta \in [1/2, 1)$ this shows that algebraic stability is not a necessary condition for the numerical solution to retain the dissipativity of the underlying system. However, we have not been able to show that A-stable is sufficient to preserve the dissipativity of (1.2.1,2.2.1) or (1.2.1,2.2.9), and neither do we establish that more general methods than the theta methods preserve the dissipativity of (1.2.1,2.2.9).

In Chapter 7 we will consider the numerical approximation of attractors and in-

variant sets of dynamical systems by Runge-Kutta methods. We again consider (1.2.1) and assume that f is at least locally Lipschitz throughout the chapter.

In Section 7.2 we show that if the underlying system has an attractor \mathcal{A} then the numerical solution defined by any Runge-Kutta method

- defines a discrete dynamical system on a neighbourhood of \mathcal{A} , with its own local attractor \mathcal{A}_h , for h sufficiently small,
- and moreover that $\text{dist}(\mathcal{A}_h, \mathcal{A}) \rightarrow 0$ as $h \rightarrow 0$.

Here $\text{dist}(\bullet, \bullet)$ is the semi-distance defined in Definition 2.3.5, and $\text{dist}(\mathcal{A}_h, \mathcal{A}) \rightarrow 0$ as $h \rightarrow 0$ implies that given any neighbourhood of \mathcal{A} , \mathcal{A}_h is contained in that neighbourhood for all sufficiently small $h > 0$. When this holds the numerical approximation is said to be *upper semicontinuous at $h = 0$* . Note that, in general \mathcal{A}_h will only be a local attractor, even when the attractor \mathcal{A} that it is approximating is globally attracting. However, if we impose additional conditions, such as requiring that the dynamical system is of the form (1.2.1, 2.2.1) and that the Runge-Kutta method used is algebraically stable, then we can ensure that the numerical attractor \mathcal{A}_h is globally attracting.

Our proof of upper semicontinuity follows the method of Hale, Lin & Raugel [29] and Temam [58]. In both of those works upper semicontinuity was proved for certain perturbations $S_\lambda(t)$ of an infinite dimensional evolution operator $S_{\lambda_0}(t)$. In [36] we proved upper semicontinuity for the numerical approximation by any Runge-Kutta method of the global attractor of a dissipative dynamical system of the form (1.2.1, 2.2.1). The result given above extends this to any attractor of any dynamical system (1.2.1) (for which f is locally Lipschitz).

A related result can be found in Kloeden and Lorenz [42]. Although they phrase their result in terms of uniformly asymptotically stable sets it can be used to show upper semicontinuity at $h = 0$ for the numerical approximation to the global attractor \mathcal{A} of a dissipative system (1.2.1). The main difference between our result and the result of [42] is that Kloeden and Lorenz make quite strong continuity and differentiability conditions on the system to ensure that the numerical solution defines a discrete dynamical system on \mathbb{R}^m and satisfies a suitable uniform local error bound; whereas we actually prove that the numerical solution defines a discrete dynamical system on a neighbourhood of \mathcal{A} and derive a global error bound under the very weak condition that f is locally Lipschitz.

Upper semicontinuity ensures that every point on the numerical attractor \mathcal{A}_h is close to a point of \mathcal{A} for h sufficiently small; but not the converse – there may be points of \mathcal{A} not approximated by the numerical attractor. To prove that \mathcal{A}_h converges to \mathcal{A} in the Hausdorff set metric as $h \rightarrow 0$ we also need to establish lower semicontinuity at $h = 0$, that is,

- $\text{dist}(\mathcal{A}, \mathcal{A}_h) \rightarrow 0$ as $h \rightarrow 0$.

This is much harder to prove than upper semicontinuity.

Hale and Raugel [30] have established lower semicontinuity results for certain perturbations of gradient systems on a Banach spaces with hyperbolic equilibria. In [36] we use the method of proof of Hale and Raugel to show lower semicontinuity at $h = 0$ of the numerical approximation, by a Runge-Kutta method, to the global attractor of a gradient system with hyperbolic equilibria and which is dissipative in the sense of (2.2.1).

In Section 7.3 we present a new proof of lower semicontinuity. Unlike the method of proof of Hale and Raugel which uses the Morse decomposition of the global attractor of a gradient system, our new approach does not require that the underlying system is in gradient form, and uses a compactness argument instead of relying on the existence of a Morse decomposition. This allows to prove that

- if the attractor of the underlying system, \mathcal{A} , is equal to the closure of the union of the unstable manifolds of its hyperbolic fixed points, then the numerical attractor \mathcal{A}_h defined by a Runge-Kutta method, as above, satisfies $\text{dist}(\mathcal{A}, \mathcal{A}_h) \rightarrow 0$ as $h \rightarrow 0$.

Since we have already proved that $\text{dist}(\mathcal{A}_h, \mathcal{A}) \rightarrow 0$ this establishes that \mathcal{A}_h converges to \mathcal{A} in the Hausdorff metric as $h \rightarrow 0$. Note that an attractor of a gradient system has the form required for our result to apply; but that we show in Example 2.3.9 that attractors of non-gradient systems can also have this form. To our knowledge, this is the first proof of lower semicontinuity for non-gradient systems.

In Section 7.4 we consider the behaviour of general numerical invariant sets and attractors \mathcal{A}_h in the limit as $h \rightarrow 0$. We introduce the concepts of $\liminf_{h \rightarrow 0} \mathcal{A}_h$ and $\limsup_{h \rightarrow 0} \mathcal{A}_h$ from set-valued analysis so that we can analyse situations where \mathcal{A}_h does not necessarily converge (to some set) in the Hausdorff metric as $h \rightarrow 0$. The \liminf and \limsup define sets which in some sense represent the smallest and largest sets of

numerically observable invariant dynamics in the limit as $h \rightarrow 0$, and \mathcal{A}_h converges in the Hausdorff metric as $h \rightarrow 0$ if and only if the \liminf and \limsup are equal; and in which case it converges to the set that they define.

For general invariant sets (not necessarily attractors) we show that

- if \mathcal{A}_h is invariant under evolution of the numerical approximation for all $h \in (0, h_0)$ and \mathcal{A}_h converges in Hausdorff metric to a compact set \mathcal{A}_0 as $h \rightarrow 0$ then \mathcal{A}_0 is invariant under evolution of the underlying system (1.2.1),

and that if \mathcal{A}_h does not converge in the Hausdorff metric then the \liminf and \limsup are both invariant under the evolution of (1.2.1). This result is unlike the other results in this chapter in that we have not made any assumptions on the dynamics of the underlying system, and have used the existence of numerically invariant sets to deduce the existence of an invariant set for the underlying system. In Chapter 4, we showed that a continuous branch of fixed points either becomes unbounded or converges to a fixed point of the underlying system as $h \rightarrow 0$; we have now extended this result to apply to the convergence of arbitrary invariant sets whether they be fixed points, periodic orbits, tori, or strange attractors.

Finally we consider the numerical approximation attractors \mathcal{A} once more, but now for general attractors which need *not* have the form assumed earlier. In this case we prove that

- if \mathcal{A}_h converges in Hausdorff metric as $h \rightarrow 0$ then it converges to a subset of \mathcal{A} which is invariant under (1.2.1) and which contains the closure of the union of the unstable manifolds of the hyperbolic fixed points of \mathcal{A} ,

whilst if \mathcal{A}_h does not converge in the Hausdorff metric then the \liminf and \limsup are both invariant subsets of \mathcal{A} which contain the closure of the union of the unstable manifolds of the hyperbolic fixed points of \mathcal{A} .

1.4 Future Research and Other Related Work

In Chapter 7 we have only considered the set convergence of attractors. This leaves two obvious avenues for future research. Firstly to establish Hausdorff convergence of invariant sets under more and more general conditions, and secondly to compare the dynamics on the numerical invariant set with the dynamics on the corresponding

invariant set of the underlying system. Also, how are the ergodic properties of the two systems related ?

It would also be desirable to extend the results in Chapters 5 and 6 to prove that the structure of the underlying system is preserved either for more methods and/or for more general systems.

In Chapters 5, 6 and 7 we have restricted attention to Runge-Kutta methods. To what extent can these results be extended to linear multistep methods ? Some results in these directions can be found in Kirchgraber [41], Eirola and Nevanlinna [15] and Hill and Suli [33].

We have considered fixed time-stepping methods throughout. Since practical codes use variable time-stepping strategies, it is essential that the dynamics of variable time-stepping methods should be analysed from a dynamical systems viewpoint. The papers of Griffiths [22], Stoffer and Nipp [52] and Stuart and Humphries [55] give some results for variable time-stepping strategies.

We have considered two classes of dynamical systems; gradient and dissipative systems. Another important class of dynamical systems which we have disregarded completely are *Hamiltonian systems*. The discovery of numerical methods which preserve the symplectic structure of the underlying Hamiltonian system was an important breakthrough in the numerical analysis of nonlinear systems. References to the numerical solution of Hamiltonian systems are too numerous to list here.

We have directly compared the discrete dynamical system defined by the numerical solution with the underlying dynamical system throughout. Another interesting approach that we have not pursued is that of *backward error analysis*. Do the discrete gradient and dissipative systems in Chapters 5 and 6 correspond to an exact solution of a perturbation of the underlying system ?

Another technique from dynamical systems theory that we have not used is that of *shadowing*. This is because we have only considered the set convergence of attractors; if the dynamics on a chaotic attractor and its numerical approximation are compared then shadowing type results are the best that we can hope to prove.

Finally we have considered the *indirect method* approach throughout. A review of some of the results from the *direct method* approach can be found in Beyn [5].

Chapter 2

Dynamical Systems

The purpose of this chapter is twofold. Firstly we will review the concepts, definitions and notation relating to dynamical systems which will be used throughout this thesis. Secondly we will introduce structural assumptions which will be imposed on the dynamical systems that we consider in later chapters, and will show what restrictions these structural assumptions impose on the possible dynamics of the system.

2.1 Introduction and Lipschitz Continuity

We begin by defining what we mean by a dynamical system on a set $U \subseteq \mathbb{R}^m$ and its evolution operator. Consider the autonomous initial value problem: find $\mathbf{y} = \mathbf{y}(t) \in U \subseteq \mathbb{R}^m$ satisfying

$$\frac{d\mathbf{y}}{dt} = \mathbf{f}(\mathbf{y}), \quad \text{and} \quad \mathbf{y}(0) = \mathbf{y}_0 \quad (2.1.1)$$

for $t \geq 0$ where $\mathbf{f}: U \rightarrow \mathbb{R}^m$.

Definition 2.1.1 The equation (2.1.1) is said to define a *dynamical system* on a set $U \subseteq \mathbb{R}^m$ if for any $\mathbf{y}_0 \in U$ there exists a unique solution of (2.1.1) with $\mathbf{y}(t) \in U$ for all $t \geq 0$. We define the *evolution operator* $S(t): U \rightarrow U$ for the dynamical system to be the operator such that $\mathbf{y}(t) = S(t)\mathbf{y}_0$. This operator has the properties that

- (i) $S(t)S(t') = S(t')S(t) = S(t+t')$ for all $t, t' \geq 0$,
- (ii) $S(0) = I$, the identity operator.

Remark In the literature these systems are often referred to as *continuous* dynamical systems as opposed to discrete dynamical systems (which will define in Section 2.5).

However the term ‘continuous dynamical system’ is also often used to indicate a dynamical system which is continuous with respect to initial data (see Definition 2.1.2 below). To avoid confusion we only use the term ‘continuous’ to refer to a system which is continuous with respect to initial data and will simply refer to systems which satisfy Definition 2.1.1 as ‘dynamical systems’.

The evolution operator $S(t)$ is merely a convenient notation for advancing the solution through time t . We will often be interested in the evolution of groups of trajectories, and for any set $E \subseteq U$ the *action* of the evolution semi-group $S(t)$ on E is defined by

$$S(t)E = \bigcup_{\mathbf{y}_0 \in E} S(t)\mathbf{y}_0.$$

We will usually be considering dynamical systems which are continuous with respect to initial data, and we now define this concept.

Definition 2.1.2 A dynamical system is said to be *continuous with respect to initial data* (or simply referred to as a *continuous dynamical system*) if given any $\mathbf{y}_0 \in U$, any $t \geq 0$ and any $\varepsilon > 0$ there exists $\delta = \delta(\mathbf{y}_0, t, \varepsilon) > 0$ such that $\|S(t)\mathbf{y}_0 - S(t)\mathbf{y}\| < \varepsilon$ for all $\mathbf{y} \in U$ such that $\|\mathbf{y} - \mathbf{y}_0\| < \delta$.

The concept of Lipschitz continuity was important in the development of existence and uniqueness theory for solutions to (2.1.1). It will no less vital in our development of theory for the numerical solution of (2.1.1); we will often assume that \mathbf{f} satisfies a Lipschitz continuity condition.

Definition 2.1.3 Given $U \subseteq \mathbb{R}^m$, $\mathbf{f}: U \rightarrow \mathbb{R}^m$ is said to be *Lipschitz* on $B \subset U$ with Lipschitz constant if L if

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in B$$

for some norm $\|\cdot\|$ on \mathbb{R}^m . If \mathbf{f} is Lipschitz on U then \mathbf{f} is said to be *globally Lipschitz*. If \mathbf{f} is Lipschitz on every bounded subset of U then \mathbf{f} is said to be *locally Lipschitz*.

Classical ordinary differential equations theory (see for example Hartman [31]) shows that if \mathbf{f} is globally Lipschitz then for any initial condition \mathbf{y}_0 the solution $\mathbf{y}(t)$ of (2.1.1) exists and is unique for all $t \in [0, t^*(\mathbf{y}_0))$, where either $t^*(\mathbf{y}_0) = \infty$ or $\mathbf{y}(t)$ leaves U at $t = t^*(\mathbf{y}_0)$. Hence if no trajectories leave U (and in particular if

$U = \mathbb{R}^m$) this is a sufficient condition to ensure that (2.1.1) defines a dynamical system on U .

More generally if \mathbf{f} is locally Lipschitz then for any bounded set $B \subseteq U$ there exists a unique solution to (2.1.1) for $0 \leq t < t^*(B)$ and any $\mathbf{y}_0 \in B$. The solution can only cease to exist if either it leaves U or it blows up, that is if $\|\mathbf{y}(t)\| \rightarrow \infty$ as $t \rightarrow t^*(B)$. If we have an a priori bound which shows that

$$\limsup_{t \rightarrow \infty} \|\mathbf{y}(t)\| \leq c(\mathbf{y}_0)$$

where $c(\mathbf{y}_0) \in \mathbb{R}$, for all $\mathbf{y}_0 \in U$ and if no trajectories leave U (in particular if $U = \mathbb{R}^m$) then a unique solution is guaranteed to exist for all $t \geq 0$ and hence (2.1.1) defines a dynamical system. We will consider several classes of problems of the form (2.1.1) where we make different structural assumptions on \mathbf{f} . These structural assumptions often imply an a priori bound on $\|\mathbf{y}_n\|$ of the type sought, and hence it will follow that these problems do define dynamical systems.

Note that the assumption that \mathbf{f} is locally Lipschitz is sufficient to ensure that if (2.1.1) defines a dynamical system then it is continuous with respect to initial data. It should also be noted that if $\mathbf{f} \in C^1(U, \mathbb{R}^m)$ then \mathbf{f} is locally Lipschitz on U . Hence any results that we prove for \mathbf{f} locally Lipschitz apply to all problems where \mathbf{f} is C^1 .

If \mathbf{f} satisfies a Lipschitz condition then, as the next result shows, the rate at which trajectories of the system may converge or diverge is bounded.

Result 2.1.4 *Suppose that (2.1.1) defines a dynamical system on \mathbb{R}^m , \mathbf{f} is Lipschitz with Lipschitz constant L on a set $B \subseteq \mathbb{R}^m$, and that $S(t)\mathbf{x}_0, S(t)\mathbf{y}_0 \in B$ for $t \in [0, t_0]$, then*

$$\|S(t)\mathbf{x}_0 - S(t)\mathbf{y}_0\| \leq e^{Lt} \|\mathbf{x}_0 - \mathbf{y}_0\| \quad (2.1.2)$$

for $t \in [0, t_0]$. If $S(-t)\mathbf{x}_0$ and $S(-t)\mathbf{y}_0$ are well defined for $-t \in [-t_0, 0]$ and $S(-t)\mathbf{x}_0, S(-t)\mathbf{y}_0 \in B$ for $-t \in [-t_0, 0]$ then

$$\|S(-t)\mathbf{x}_0 - S(-t)\mathbf{y}_0\| \leq e^{Lt} \|\mathbf{x}_0 - \mathbf{y}_0\| \quad (2.1.3)$$

for $-t \in [-t_0, 0]$.

Proof. First we derive (2.1.2). Using the Cauchy-Schwarz inequality and Lipschitz

continuity implies

$$\begin{aligned}\frac{d}{dt}\|\mathbf{x}(t) - \mathbf{y}(t)\|^2 &= 2\langle \mathbf{x}(t) - \mathbf{y}(t), \mathbf{f}(\mathbf{x}(t)) - \mathbf{f}(\mathbf{y}(t)) \rangle \\ &\leq 2\|\mathbf{x}(t) - \mathbf{y}(t)\| \|\mathbf{f}(\mathbf{x}(t)) - \mathbf{f}(\mathbf{y}(t))\| \\ &\leq 2L\|\mathbf{x}(t) - \mathbf{y}(t)\|^2.\end{aligned}$$

for $t \in [0, t_0]$. Thus for $0 \leq s \leq t \leq t_0$

$$\begin{aligned}e^{-2Ls}\frac{d}{ds}\|\mathbf{x}(s) - \mathbf{y}(s)\|^2 - e^{-2Ls}2L\|\mathbf{x}(s) - \mathbf{y}(s)\|^2 &\leq 0 \\ \frac{d}{ds}\left[e^{-2Ls}\|\mathbf{x}(s) - \mathbf{y}(s)\|^2\right] &\leq 0 \\ \left[e^{-2Ls}\|\mathbf{x}(s) - \mathbf{y}(s)\|^2\right]_{s=0}^t &\leq 0\end{aligned}$$

and taking square roots and rearranging implies (2.1.2). To prove (2.1.3) note that

$$\frac{d}{dt}\|\mathbf{x}(t) - \mathbf{y}(t)\|^2 \geq -2\|\mathbf{x}(t) - \mathbf{y}(t)\| \|\mathbf{f}(\mathbf{x}(t)) - \mathbf{f}(\mathbf{y}(t))\|$$

and proceed as in proof of (2.1.2). ■

The following lemma will be used in Chapter 3 to derive error bounds for the numerical solution of (2.1.1) by Runge-Kutta methods when \mathbf{f} is Lipschitz continuous, but not necessarily differentiable.

Lemma 2.1.5 *Suppose that (2.1.1) defines a dynamical system on \mathbb{R}^m , \mathbf{f} is Lipschitz with Lipschitz constant L on a set $B \subseteq \mathbb{R}^m$, and that $S(t)\mathbf{y}_0 \in B$ for $t \in [0, t_0]$, then*

$$\|S(t)\mathbf{y}_0 - \mathbf{y}_0\| \leq \frac{1}{L}\|\mathbf{f}(\mathbf{y}_0)\|[e^{Lt} - 1] \quad (2.1.4)$$

for $t \in [0, t_0]$.

Proof. Let $\mathbf{x}(s) = \mathbf{y}(s) - \mathbf{y}_0$, where $\mathbf{y}(s) = S(s)\mathbf{y}_0$. First we show that

$$\frac{d}{ds}\|\mathbf{x}(s)\| \leq \|\mathbf{f}(\mathbf{y}(s))\|. \quad (2.1.5)$$

Since $\frac{d\mathbf{x}}{ds} = \frac{d\mathbf{y}}{ds} = \mathbf{f}(\mathbf{y}(s))$ we see that

$$2\|\mathbf{x}(s)\|\frac{d}{ds}\|\mathbf{x}(s)\| = \frac{d}{ds}\|\mathbf{x}(s)\|^2 = 2\langle \mathbf{x}(s), \mathbf{f}(\mathbf{y}(s)) \rangle \leq 2\|\mathbf{x}(s)\|\|\mathbf{f}(\mathbf{y}(s))\|$$

and (2.1.5) follows for $\mathbf{x}(s) \neq 0$. Now, by continuity of \mathbf{f} , (2.1.5) must also hold for $\mathbf{x}(s) = 0$. Now for $0 \leq s \leq t \leq t_0$

$$\begin{aligned} \frac{d}{ds} \|\mathbf{x}(s)\| &\leq \|\mathbf{f}(\mathbf{y}(s)) - \mathbf{f}(\mathbf{y}_0) + \mathbf{f}(\mathbf{y}_0)\| \\ &\leq \|\mathbf{f}(\mathbf{y}(s)) - \mathbf{f}(\mathbf{y}_0)\| + \|\mathbf{f}(\mathbf{y}_0)\| \\ &\leq L\|\mathbf{y}(s) - \mathbf{y}_0\| + \|\mathbf{f}(\mathbf{y}_0)\| \\ &= L\|\mathbf{x}(s)\| + \|\mathbf{f}(\mathbf{y}_0)\| \end{aligned}$$

and hence

$$\begin{aligned} e^{-Ls} \frac{d}{ds} \|\mathbf{x}(s)\| - e^{-Ls} L \|\mathbf{x}(s)\| &\leq e^{-Ls} \|\mathbf{f}(\mathbf{y}_0)\| \\ \frac{d}{ds} [e^{-Ls} \|\mathbf{x}(s)\|] &\leq e^{-Ls} \|\mathbf{f}(\mathbf{y}_0)\| \\ [e^{-Ls} \|\mathbf{x}(s)\|]_{s=0}^t &\leq \|\mathbf{f}(\mathbf{y}_0)\| \int_0^t e^{-Ls} ds \\ e^{-Lt} \|\mathbf{x}(t)\| &\leq \frac{1}{L} \|\mathbf{f}(\mathbf{y}_0)\| [-e^{-Ls}]_{s=0}^t \end{aligned}$$

as required. ■

Instead of assuming Lipschitz continuity we will sometimes consider (2.1.1) under the assumption that \mathbf{f} satisfies a one-sided Lipschitz condition.

Definition 2.1.6 The function $\mathbf{f}: U \rightarrow \mathbb{R}^m$ is said to satisfy a *one-sided Lipschitz condition* on U if

$$\langle \mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle \leq c \|\mathbf{u} - \mathbf{v}\|^2 \quad \forall \mathbf{u}, \mathbf{v} \in U \quad (2.1.6)$$

for some $c \in \mathbb{R}$.

Using the Cauchy-Schwarz inequality it is easy to show that if \mathbf{f} is Lipschitz on U then it satisfies a one-sided Lipschitz condition on U , and hence every globally Lipschitz function satisfies a one-sided Lipschitz condition. Moreover, the one-sided Lipschitz constant c is less than or equal to the global Lipschitz constant L .

The problem (2.1.1, 2.1.6) has been studied extensively in the “stiff” numerical analysis literature, and this work is reviewed in Dekker and Verwer [13] and Hairer and

Wanner [27]. The theory in [13, 27] is for nonautonomous systems which satisfy

$$\langle \mathbf{f}(t, \mathbf{u}) - \mathbf{f}(t, \mathbf{v}), \mathbf{u} - \mathbf{v} \rangle \leq c \|\mathbf{u} - \mathbf{v}\|^2 \quad \forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^m$$

but we will restrict our attention to autonomous systems. The following result shows that the one-sided Lipschitz condition imposes a bound on the rate at which trajectories of the system may diverge.

Result 2.1.7 *If $\mathbf{u}(t)$, $\mathbf{v}(t)$ are two solutions of (2.1.1, 2.1.6) with initial conditions \mathbf{u}_0 and \mathbf{v}_0 respectively then*

$$\|\mathbf{u}(t) - \mathbf{v}(t)\| \leq e^{ct} \|\mathbf{u}_0 - \mathbf{v}_0\|. \quad (2.1.7)$$

Proof. Consider

$$\begin{aligned} \frac{d}{dt} \|\mathbf{u} - \mathbf{v}\|^2 &= 2\langle \mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle \\ &\leq 2c \|\mathbf{u} - \mathbf{v}\|^2, \end{aligned}$$

hence

$$\begin{aligned} e^{2ct} \frac{d}{dt} \|\mathbf{u} - \mathbf{v}\|^2 + 2ce^{2ct} \|\mathbf{u} - \mathbf{v}\|^2 &\leq 0 \\ \frac{d}{dt} [e^{2ct} \|\mathbf{u}(t) - \mathbf{v}(t)\|^2] &\leq 0 \end{aligned}$$

and integrating implies that

$$e^{2ct} \|\mathbf{u}(t) - \mathbf{v}(t)\|^2 - \|\mathbf{u}_0 - \mathbf{v}_0\|^2 \leq 0.$$

Rearranging and taking square roots gives (2.1.7). ■

Remark Unlike Lipschitz continuity the one-sided Lipschitz condition does not impose any bound on the rate at which trajectories may converge.

If $c < 0$ then the system (2.1.1, 2.1.6) is said to be *exponentially contractive*, because (2.1.7) then implies exponential convergence of trajectories.

If $c = 0$ then

$$\langle \mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle \leq 0 \quad (2.1.8)$$

for all $u, v \in \mathbb{R}^m$ and for all $t \geq 0$. Equation (2.1.7) then implies that $\|u(t) - v(t)\|$ is nonincreasing in t . For this reason such systems are often referred to as *contractive* in the numerical analysis literature, where they have been used extensively in nonlinear stability definitions for numerical methods. However, as Theorem 2.1.8 shows, from a dynamical systems perspective, the contractive problems (2.1.1, 2.1.8) are a somewhat unnatural class of systems to consider. Generically fixed points of dynamical systems are hyperbolic and isolated, but the following theorem shows that this is not the case for (2.1.1, 2.1.8).

Theorem 2.1.8 *Suppose that (2.1.1, 2.1.8) defines a continuous dynamical system on \mathbb{R}^m and let $\mathcal{E} = \{x: f(x) = 0\}$, the set of fixed points of (2.1.1, 2.1.8). Then \mathcal{E} is a closed convex set.*

Proof. Since f is continuous \mathcal{E} must be closed. To prove that \mathcal{E} is convex it is sufficient to show that any convex combination of two zeros of f is also a zero of f . So suppose $f(x) = 0$, $f(y) = 0$, $x \neq y$ and $\lambda \in (0, 1)$. Let $z = \lambda x + (1 - \lambda)y$ and suppose also that $f(z) \neq 0$. Let $w = z + \delta f(z)$ then

$$\langle f(w) - f(x), w - x \rangle \leq 0$$

implies

$$\begin{aligned} \langle f(w), \delta f(z) + z - x \rangle &\leq 0 \\ \langle f(w), \delta f(z) + (\lambda - 1)(x - y) \rangle &\leq 0 \\ \langle f(w), \delta f(z) \rangle &\leq (1 - \lambda) \langle f(w), x - y \rangle. \end{aligned} \quad (2.1.9)$$

Also

$$\langle f(w) - f(y), w - y \rangle \leq 0$$

implies

$$\begin{aligned} \langle f(w), \delta f(z) + z - y \rangle &\leq 0 \\ \langle f(w), \delta f(z) + \lambda(x - y) \rangle &\leq 0 \\ \langle f(w), \delta f(z) \rangle &\leq -\lambda \langle f(w), x - y \rangle. \end{aligned} \quad (2.1.10)$$

Notice that $(1 - \lambda)$ and $-\lambda$ have opposite signs, hence (2.1.9) and (2.1.10) imply that

$$\langle f(w), \delta f(z) \rangle \leq 0,$$

and since the sign of δ is arbitrary it follows that

$$\langle f(z + \delta f(z)), f(z) \rangle = 0,$$

and allowing $\delta \rightarrow 0$ we obtain $\|f(z)\|^2 = 0$, and convexity of \mathcal{E} follows. ■

We would like to also show that $\text{dist}(y(t), \mathcal{E}) = \inf_{x \in \mathcal{E}} \|y(t) - x\| \rightarrow 0$ as $t \rightarrow \infty$ for any y_0 . In Result 3.1 of [56] we show this in the case where inequality is strict in (2.1.8) for each $u \notin \mathcal{E}$ and each $v \in \mathcal{E}$. The following example shows that this result fails without this extra condition.

Example 2.1.9 For $m \geq 2$ we can construct an example of a system (2.1.1,2.1.8) where all trajectories rotate about \mathcal{E} and thus $\text{dist}(y(t), \mathcal{E})$ is fixed. Let y_i be the i -th coordinate of y . Then consider

$$\begin{aligned} \dot{y}_1 &= -y_2 \\ \dot{y}_2 &= y_1 \\ \dot{y}_i &= 0 \quad \text{for } i = 3, \dots, m. \end{aligned}$$

Then $\mathcal{E} = \{y: y_1 = y_2 = 0\}$ and $\|u(t) - v(t)\|$ is constant, where $u(t)$ and $v(t)$ are two solutions of this system with different initial conditions. □

Stuart (private communication) has shown that the if $c < 0$ then (2.1.1,2.1.6) must have a unique globally attracting fixed point. Because of the essentially trivial nature of the dynamics of these problems, we will not consider their numerical solution here, but this and related problems are considered in [35]. It should be noted that although systems of the form (2.1.1,2.1.8) do not have interesting dynamics, these systems have received a great deal of attention in the numerical analysis literature, and the results concerning them and their numerical solutions together with the mathematics used to prove these results has been very important in the development of numerical analysis for nonlinear ordinary differential equations.

2.2 Dissipative Dynamical Systems

Instead of considering systems for which all trajectories are asymptotic to a unique fixed point, a natural generalization is to consider systems for which the asymptotic behaviour is confined to some bounded set, but where no restrictions are imposed on the possible dynamics within this set. Such systems are said to be dissipative, and we will consider their numerical solution in Chapter 6.

Definition 2.2.1 If (2.1.1) defines a dynamical system on $U \subseteq \mathbb{R}^m$ then this system is said to be *dissipative* if there is a bounded set B with the property that, for any bounded set $E \subseteq U$, there exists $t^* = t^*(E) \geq 0$ such that $S(t)E \subseteq B$ for all $t > t^*$. The set B is called an *absorbing set*.

Remark Hale [28] notes that for a continuous dynamical system on a locally compact space (such as \mathbb{R}^m) to show dissipativity it is sufficient to show that for any initial condition $y_0 \in U$ there exists $t^*(y_0) \geq 0$ such that $S(t)y_0 \in B$ for $t > t^*$. This will simplify proofs of dissipativity.

Note that absorbing sets are not unique, since if B is an absorbing set then any bounded set B' such that $B \subset B'$ will also be an absorbing set.

Remark Sometimes a system is said to be dissipative if the divergence of the flow is negative, but this is not equivalent to the definition given above, and when we refer to a system as being dissipative we will always mean dissipative in the sense of Definition 2.2.1.

2.2.1 Structural Assumptions Inducing Dissipativity

Often dissipativity is a direct consequence of a structural condition which f satisfies. In this section we will consider two such structural assumptions.

First consider (2.1.1) under the assumptions that f is locally Lipschitz and that there exist constants $\alpha \geq 0$ and $\beta > 0$ such that

$$\langle f(y), y \rangle \leq \alpha - \beta \|y\|^2 \quad \text{for all } y \in \mathbb{R}^m \quad (2.2.1)$$

where the norm in (2.2.1) is the norm induced by the inner product. Under these assumptions (2.1.1) defines a dissipative dynamical system, as we will now show.

Theorem 2.2.2 *If $f: \mathbb{R}^m \rightarrow \mathbb{R}^m$ is locally Lipschitz then (2.1.1, 2.2.1) defines a dynamical system on \mathbb{R}^m and for any $\varepsilon > 0$ there exists $t^* = t^*(\mathbf{y}_0, \varepsilon)$ such that for all $t > t^*$*

$$\|\mathbf{y}(t)\|^2 < \frac{\alpha}{\beta} + \varepsilon. \quad (2.2.2)$$

Hence the dynamical system is dissipative, and the open ball $B = B(0, \sqrt{\alpha/\beta} + \varepsilon)$ is an absorbing set for any $\varepsilon > 0$.

Proof. First we establish an a priori bound on the solution $\mathbf{y}(t)$. Note that

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \|\mathbf{y}(t)\|^2 &= \langle f(\mathbf{y}(t)), \mathbf{y}(t) \rangle \\ &\leq \alpha - \beta \|\mathbf{y}(t)\|^2. \end{aligned} \quad (2.2.3)$$

Hence

$$\begin{aligned} e^{2\beta t} \frac{d}{dt} \|\mathbf{y}(t)\|^2 + 2\beta e^{2\beta t} \|\mathbf{y}(t)\|^2 &\leq 2\alpha e^{2\beta t} \\ \frac{d}{dt} [e^{2\beta t} \|\mathbf{y}(t)\|^2] &\leq 2\alpha e^{2\beta t} \\ e^{2\beta t} \|\mathbf{y}(t)\|^2 - \|\mathbf{y}_0\|^2 &\leq \frac{\alpha}{\beta} [e^{2\beta t} - 1] \\ \|\mathbf{y}(t)\|^2 &\leq \frac{\alpha}{\beta} + e^{-2\beta t} \left[\|\mathbf{y}_0\|^2 - \frac{\alpha}{\beta} \right]. \end{aligned} \quad (2.2.4)$$

Thus it follows that

$$\|\mathbf{y}(t)\| \leq \max \left(\|\mathbf{y}_0\|, \sqrt{\frac{\alpha}{\beta}} \right)$$

for $t \geq 0$. Hence the solution of (2.1.1) cannot blow up and since f is locally Lipschitz it follows from the remarks on page 27 that (2.1.1, 2.2.1) defines a dynamical system on \mathbb{R}^m .

The bound (2.2.2) follows from (2.2.4) and this implies that (2.1.1, 2.2.1) is dissipative, with B an absorbing set. ■

Remark Equation (2.2.3) implies that

$$\limsup_{\|\mathbf{y}\| \rightarrow \infty} \frac{\frac{d}{dt} \|\mathbf{y}\|}{\|\mathbf{y}\|} \leq -\beta$$

so that the decay from ‘infinity’ is at least linear for systems of the form (2.1.1, 2.2.1).

Although the asymptotic behaviour of a dissipative system is confined to a bounded

set, we emphasise the fact that the dynamics within this set may be extremely complicated. To stress this we present an example which shows that the Lorenz equations define a dissipative dynamical system which satisfies (2.2.1) after translation of the coordinate system.

Example 2.2.3 Consider the Lorenz system of ordinary differential equations in \mathbb{R}^3 defined by

$$\left. \begin{aligned} \dot{x} &= \sigma(y - x) \\ \dot{y} &= rx - y - xz \\ \dot{z} &= xy - bz \end{aligned} \right\} \quad (2.2.5)$$

where $x(t)$, $y(t)$ and $z(t)$ are to be found for $t \geq 0$ and σ , r and b are positive parameters. This system was first introduced by Lorenz [44] and arises as a finite dimensional spectral truncation of equations governing Rayleigh-Bénard convection. The parameters are often taken to be $\sigma = 10$, $r = 25$ and $b = 8/3$. For these values the system appears to display chaotic dynamics and to possess a strange attractor, and as an early example of such a system has been important in the development of chaos theory.

If we write $\mathbf{y} = (x, y, z)^T$ and $\mathbf{f}(\mathbf{y}) = (\sigma(y - x), rx - y - xz, xy - bz)^T$ then we see that this system is of the form (2.1.1) with $\mathbf{f} \in C^\infty(\mathbb{R}^3, \mathbb{R}^3)$ and hence locally Lipschitz. To show that the Lorenz equations define a dissipative system we translate the coordinate system by $z \mapsto z - r - \sigma$, obtaining

$$\left. \begin{aligned} \dot{x} &= \sigma(y - x) \\ \dot{y} &= -\sigma x - y - xz \\ \dot{z} &= xy - bz - b(r + \sigma). \end{aligned} \right\} \quad (2.2.6)$$

Now defining \mathbf{y} as above and \mathbf{f} by $\mathbf{f}(\mathbf{y}) = (\sigma(y - x), -\sigma x - y - xz, xy - bz - b(r + \sigma))^T$ and using the Euclidean inner product we obtain

$$\langle \mathbf{f}(\mathbf{y}), \mathbf{y} \rangle = -\sigma x^2 - y^2 - bz^2 - bz(r + \sigma). \quad (2.2.7)$$

Temam [58] shows that if $b > 1$ then (2.2.7) implies that

$$\langle \mathbf{f}(\mathbf{y}), \mathbf{y} \rangle \leq -\sigma x^2 - y^2 - z^2 + \frac{b^2(r + \sigma)^2}{4(b - 1)} \quad (2.2.8)$$

and hence (2.2.1) is satisfied with

$$\begin{aligned}\alpha &= \frac{b^2(r + \sigma)^2}{4(b - 1)} \\ \beta &= \min(1, \sigma).\end{aligned}$$

Hence the translated Lorenz equations (2.2.6) are dissipative, and reversing the translation it follows that the Lorenz equations (2.2.5) are themselves dissipative. \square

We now consider a generalization of the condition (2.2.1). Notice that if \mathbf{f} satisfies (2.2.1) then

$$\langle \mathbf{f}(\mathbf{y}), \mathbf{y} \rangle < 0 \quad \text{for } \|\mathbf{y}\| > R \quad (2.2.9)$$

for any $R \geq \sqrt{\alpha/\beta}$. The theorem below shows that if \mathbf{f} satisfies (2.2.9) then (2.1.1) defines a dissipative dynamical system and hence that (2.2.9) defines a generalization of the class of dissipative problems defined by (2.2.1).

Theorem 2.2.4 *If $\mathbf{f}: \mathbb{R}^m \rightarrow \mathbb{R}^m$ is locally Lipschitz then (2.1.1, 2.2.9) defines a dissipative dynamical system, and the open ball $B(0, R + \varepsilon)$ is an absorbing set for any $\varepsilon > 0$.*

Proof. Given $\varepsilon > 0$ let B denote the open ball $B(0, R + \varepsilon)$. Now for $\mathbf{y} \in \mathbb{R}^m \setminus B$

$$\begin{aligned}\frac{d}{dt} \|\mathbf{y}(t)\|^2 &= 2\langle \mathbf{f}(\mathbf{y}(t)), \mathbf{y}(t) \rangle \\ &< 0,\end{aligned}$$

hence given any \mathbf{y}_0 it follows that $\|\mathbf{y}(t)\| < \|\mathbf{y}_0\|$ for all $t > 0$ and solutions cannot blow up. Therefore \mathbf{f} locally Lipschitz is sufficient to ensure that (2.1.1, 2.2.9) defines a dynamical system. Now given any bounded set E , let $r = \sup_{\mathbf{y} \in E} \|\mathbf{y}\|$ and $E^* = \{\mathbf{y}: \|\mathbf{y}\| \leq r\}$. Now note that for $\mathbf{y} \in E^* \setminus B$

$$\frac{d}{dt} \|\mathbf{y}\|^2 < 0.$$

Hence E^* and B are forward invariant, that is $S(t)E^* \subseteq E^*$ and $S(t)B \subseteq B$ for all $t \geq 0$. Note also that $E^* \setminus B$ is a compact set and $\langle \mathbf{f}(\mathbf{y}), \mathbf{y} \rangle < 0$ on $E^* \setminus B$ hence by continuity of \mathbf{f} , there exists $\delta > 0$ such that

$$\frac{d}{dt} \|\mathbf{y}\|^2 \leq -\delta \quad \text{on } E^* \setminus B. \quad (2.2.10)$$

Thus for any $\mathbf{y}_0 \in E$, it follows that $\mathbf{y}(t) \in B$ for $t > [r^2 - (R + \varepsilon)^2]/\delta$ and hence this system is dissipative. ■

2.2.2 Absence of One-sided Lipschitz Condition

Since all the trajectories of a dissipative system enter a bounded absorbing set it would seem natural that the rate of divergence of trajectories would be bounded, and that dissipative systems would satisfy one-sided Lipschitz conditions. However in this section we will present examples which show that this is false. Dissipative systems in general and the Lorenz equations in particular do not satisfy a one-sided Lipschitz condition on \mathbb{R}^m , and so we will not assume that such a condition when we consider the numerical solution of these systems in Chapter 6. We begin with a simple two-dimensional example.

Example 2.2.5 Consider the following two-dimensional problem in polar coordinates

$$\begin{aligned}\dot{r} &= -r \\ \dot{\theta} &= -r \cos \theta.\end{aligned}$$

Converting to Cartesian coordinates gives

$$\begin{aligned}\dot{x} &= -x + xy \\ \dot{y} &= -y - x^2.\end{aligned}$$

We will show that this system satisfies (2.2.1) and hence is dissipative, but that for the Euclidean inner product the system does not satisfy a one-sided Lipschitz condition. To show dissipativity let $\mathbf{u} = (x, y)^T$ and $\mathbf{f}(\mathbf{u}) = (-x + xy, -y - x^2)^T$ then we may rewrite the system as $\dot{\mathbf{u}} = \mathbf{f}(\mathbf{u})$ where, using the Euclidean inner product,

$$\begin{aligned}\langle \mathbf{f}(\mathbf{u}), \mathbf{u} \rangle &= -x^2(1 - y) - y(y + x^2) \\ &= -x^2 - y^2 \\ &= -\|\mathbf{u}\|^2.\end{aligned}$$

Thus (2.2.1) is satisfied with $\alpha = 0$ and $\beta = 1$, and hence the system is dissipative. Indeed, since $\alpha = 0$, Theorem 2.2.2 implies that $\|\mathbf{u}(t)\| \rightarrow 0$ as $t \rightarrow \infty$ for any initial

condition \mathbf{u}_0 . However the system does not satisfy a one-sided Lipschitz condition, as we will now show. Let $\mathbf{v} = (x', y')$ then

$$\langle \mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle = -(x - x')^2 + (x - x')(xy - x'y') - (y - y')^2 + (y - y')(x'^2 - x^2)$$

Now suppose that a one-sided Lipschitz condition (2.1.6) holds and let

$$\mathbf{u} = [c_2, c_1]^T \quad \mathbf{v} = [c_1, c_2]^T$$

where we will specify the constants c_1 and c_2 below in order to obtain a contradiction.

Notice that $\|\mathbf{u} - \mathbf{v}\|^2 = 2(c_2 - c_1)^2$ and observe that

$$\begin{aligned} \langle \mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle &= -2(c_2 - c_1)^2 + (c_2 - c_1)(c_2^2 - c_1^2) \\ &= -2(c_2 - c_1)^2 \left[1 - \frac{1}{2}(c_1 + c_2) \right] \\ &= \left[\frac{1}{2}(c_1 + c_2) - 1 \right] \|\mathbf{u} - \mathbf{v}\|^2. \end{aligned}$$

Now choosing $c_1 + c_2 > 2(c + 1)$ contradicts (2.1.6). Thus this system does not satisfy a one-sided Lipschitz condition for the Euclidean inner product for any $c > 0$, even though this system is dissipative, satisfies (2.2.1), and the origin is globally attracting.

□

The above example is not an isolated case. In [36] it is shown that the Lorenz equations (2.2.5) do not satisfy a one-sided Lipschitz condition under the Euclidean inner product. Here we will extend this result to general inner products. A general inner product on \mathbb{R}^m is defined by

$$\langle \mathbf{u}, \mathbf{v} \rangle_A = \langle \mathbf{u}, A\mathbf{v} \rangle_E$$

where A is a positive definite symmetric matrix and $\langle \bullet, \bullet \rangle_E$ represents the Euclidean inner product. We now prove that the Lorenz equations do not satisfy a one-sided Lipschitz condition for any inner product.

Theorem 2.2.6 *The Lorenz equations (2.2.5) do not satisfy a one-sided Lipschitz condition (2.1.6) for any $c > 0$ for any inner product on \mathbb{R}^3 .*

Proof. Consider a general symmetric positive definite 3×3 matrix A .

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{pmatrix}$$

Now suppose that a one-sided Lipschitz condition

$$\langle f(u) - f(v), u - v \rangle_A \leq c \|u - v\|_A^2 \quad (2.2.11)$$

holds for some $c > 0$, and derive a contradiction. Let

$$u = [c_2, c_1, c_2]^T \quad v = [c_1, c_2, c_1]^T$$

where we will specify the constants c_1 and c_2 below, in order to obtain a contradiction.

Notice that

$$\begin{aligned} \|u - v\|_A^2 &= (a_{11} + a_{22} + a_{33} - 2a_{12} + 2a_{13} - 2a_{23})(c_2 - c_1)^2 \\ &= k_0(c_2 - c_1)^2 \end{aligned}$$

and observe that

$$\begin{aligned} \langle f(u) - f(v), u - v \rangle_A &= \begin{bmatrix} 2\sigma(c_1 - c_2) \\ (1+r)(c_2 - c_1) - (c_2^2 - c_1^2) \\ b(c_1 - c_2) \end{bmatrix}^T \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{bmatrix} \begin{bmatrix} c_2 - c_1 \\ c_1 - c_2 \\ c_2 - c_1 \end{bmatrix} \\ &= (c_2 - c_1)^2 [1, -1, 1] A \begin{bmatrix} -2\sigma \\ 1+r \\ -b \end{bmatrix} - (c_2 - c_1)^2 [1, -1, 1] A \begin{bmatrix} 0 \\ c_2 + c_1 \\ 0 \end{bmatrix} \\ &= k_1(c_2 - c_1)^2 + (a_{22} - a_{12} - a_{23})(c_2 + c_1)(c_2 - c_1)^2 \\ &= \frac{1}{k_0} [k_1 + (a_{22} - a_{12} - a_{23})(c_2 + c_1)] \|u - v\|_A^2 \end{aligned}$$

Now if $a_{22} - a_{12} - a_{23} \neq 0$ then choosing $c_1 + c_2 > \frac{ck_0 - k_1}{a_{22} - a_{12} - a_{23}}$ contradicts (2.2.11), hence

$$a_{22} - a_{12} - a_{23} = 0. \quad (2.2.12)$$

Now let

$$\mathbf{u} = [c_2, 0, c_2]^T \quad \mathbf{v} = [c_1, 0, c_1]^T$$

and noting that

$$\begin{aligned} \|\mathbf{u} - \mathbf{v}\|_A^2 &= (a_{11} + 2a_{13} - 2a_{33})(c_2 - c_1)^2 \\ &= k_2(c_2 - c_1)^2 \end{aligned}$$

it follows that

$$\begin{aligned} \langle \mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle_A &= \begin{bmatrix} \sigma(c_1 - c_2) \\ r(c_2 - c_1) - (c_2^2 - c_1^2) \\ b(c_1 - c_2) \end{bmatrix}^T \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{bmatrix} \begin{bmatrix} c_2 - c_1 \\ 0 \\ c_2 - c_1 \end{bmatrix} \\ &= (c_2 - c_1)^2 [1, 0, 1] A \begin{bmatrix} -\sigma \\ r \\ -b \end{bmatrix} - (c_2 - c_1)^2 [1, 0, 1] A \begin{bmatrix} 0 \\ c_2 + c_1 \\ 0 \end{bmatrix} \\ &= k_3(c_2 - c_1)^2 - (a_{12} + a_{23})(c_2 + c_1)(c_2 - c_1)^2 \\ &= \frac{1}{k_2} [k_3 - (a_{12} + a_{23})(c_2 + c_1)] \|\mathbf{u} - \mathbf{v}\|_A^2 \end{aligned}$$

Now if $a_{12} + a_{23} \neq 0$ then choosing $c_1 + c_2 > -\frac{ck_2 - k_3}{a_{12} + a_{23}}$ contradicts (2.2.11), hence

$$a_{12} + a_{23} = 0 \tag{2.2.13}$$

but (2.2.12) and (2.2.13) imply that $a_{22} = 0$ and this contradicts the positive definiteness of A , hence the one-sided Lipschitz condition (2.2.11) does not hold for any $c > 0$, as required. ■

In Example 2.2.3 we considered a translation of the Lorenz equations in order to show that they defined a dissipative dynamical system of the form (2.1.1, 2.2.1). We show here that if \mathbf{f} does not satisfy a one-sided Lipschitz condition then no translation of the system will satisfy a one-sided Lipschitz condition. To see this note that if \mathbf{f} satisfies (2.1.6) and

$$\mathbf{g}(\mathbf{y}) = \mathbf{f}(\mathbf{y} - \boldsymbol{\alpha})$$

for an arbitrary translation α , then

$$\begin{aligned}\langle g(u) - g(v), u - v \rangle &= \langle f(u - \alpha) - g(v - \alpha), (u - \alpha) - (v - \alpha) \rangle \\ &\leq c \|(u - \alpha) - (v - \alpha)\|^2 \\ &= c\|u - v\|^2.\end{aligned}$$

Hence f satisfies a one-sided Lipschitz condition if and only if every translation of f satisfies a one-sided Lipschitz condition.

Although dissipative systems of the form (2.1.1, 2.2.1) do not in general satisfy a one-sided Lipschitz condition, some systems of this form do satisfy such a condition. For the final example in this section we exhibit a system of the form (2.1.1, 2.2.1) where f satisfies a one-sided Lipschitz condition (but is not globally Lipschitz).

Example 2.2.7 Consider the two-dimensional problem, in polar coordinates

$$\left. \begin{aligned}\dot{r} &= r(1 - r^2) \\ \dot{\theta} &= 1.\end{aligned} \right\} \quad (2.2.14)$$

Converting to Cartesian coordinates we obtain

$$\left. \begin{aligned}\dot{x} &= x - y - x(x^2 + y^2) \\ \dot{y} &= x + y - y(x^2 + y^2)\end{aligned} \right\} \quad (2.2.15)$$

Let $u = (x, y)^T$ and $f(u) = (x - y - x(x^2 + y^2), x + y - y(x^2 + y^2))^T$ then for the Euclidean inner product $\|u\| = r$ and it follows that

$$\begin{aligned}\langle f(u), u \rangle &= \frac{1}{2} \frac{d}{dt} \|u\|^2 \\ &= \frac{1}{2} \frac{d}{dt} (r^2) \\ &= r\dot{r} \\ &= r^2(1 - r^2) \\ &= \|u\|^2 - \|u\|^4 \\ &\leq 1 - \|u\|^2.\end{aligned}$$

Thus (2.2.1) is satisfied with $\alpha = \beta = 1$ and this is a dissipative system. We will now show that this system satisfies a one-sided Lipschitz condition (2.1.6). Let $v = (x', y')^T$

then

$$\begin{aligned}\langle \mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle &= (x - x')^2 + (y - y')^2 - (x^2 + y^2)^2 - (x'^2 + y'^2)^2 \\ &\quad + (xx' + yy')[(x^2 + y^2) + (x'^2 + y'^2)].\end{aligned}$$

Note that

$$xx' = \frac{1}{2}(x^2 + x'^2) - \frac{1}{2}(x - x')^2.$$

Hence

$$\begin{aligned}xx' + yy' &= \frac{1}{2}(x^2 + y^2 + x'^2 + y'^2) - \frac{1}{2}(x - x')^2 - \frac{1}{2}(y - y')^2 \\ &= \frac{1}{2}[\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - \|\mathbf{u} - \mathbf{v}\|^2]\end{aligned}$$

and it follows that

$$\begin{aligned}\langle \mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle &= \|\mathbf{u} - \mathbf{v}\|^2 - \|\mathbf{u}\|^4 + \|\mathbf{v}\|^4 \\ &\quad + \frac{1}{2}[\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - \|\mathbf{u} - \mathbf{v}\|^2][\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2] \\ &= \|\mathbf{u} - \mathbf{v}\|^2 \left[1 - \frac{1}{2}(\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2)\right] + \frac{1}{2}[\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2]^2 - \|\mathbf{u}\|^4 + \|\mathbf{v}\|^4 \\ &= \|\mathbf{u} - \mathbf{v}\|^2 \left[1 - \frac{1}{2}(\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2)\right] - \frac{1}{2}[\|\mathbf{u}\|^2 - \|\mathbf{v}\|^2]^2 \\ &\leq \|\mathbf{u} - \mathbf{v}\|^2.\end{aligned}$$

Thus \mathbf{f} satisfies the one-sided Lipschitz condition (2.1.6) with $c = 1$. \square

2.3 Invariant Sets and Attractors

A dissipative dynamical system always has a global attractor. To enable us to study attractors we must first make some definitions.

Definition 2.3.1 For the dynamical system (2.1.1) the *positive orbit* through \mathbf{x} is defined by

$$\Gamma^+(\mathbf{x}) = \bigcup_{t \geq 0} S(t)\mathbf{x}.$$

A *negative orbit* through \mathbf{x} , $\Gamma^-(\mathbf{x})$, if it exists is a set $\{\mathbf{x}(t) : t \leq 0\}$ such that

- (i) $S(-t)\mathbf{x}(t) = \mathbf{x}_0$ for all $t \leq 0$, and,

(ii) $S(t_2 - t_1)x(t_1) = x(t_2)$ for all $t_1 \leq t_2 \leq 0$.

We call $\Gamma(x)$, where

$$\Gamma(x) = \Gamma^+(x) \cup \Gamma^-(x)$$

a *complete orbit* through x .

Note that a negative orbit need not exist, since solutions may blow up in finite negative time. However if f is locally Lipschitz then the negative orbit $\Gamma^-(x)$ is unique, if it exists.

Definition 2.3.2 A set $E \subset U$ is *forward invariant* under the evolution operator S if $\Gamma^+(x) \subseteq E$ for all x in E . $E \subset U$ is said to be *backward invariant* under S if for each $x \in E$ there exists a negative orbit $\Gamma^-(x)$ such that $\Gamma^-(x) \subseteq E$. E is *invariant* under S if E is both forward and backward invariant under S .

So E is forward (resp. backward) invariant under S if $S(t)x \in E$ for all $x \in E$ and all $t \geq 0$ (resp. $t \leq 0$). Note that if E is invariant then there exists a complete orbit through each point of E . Where it is clear which evolution operator is under consideration we will sometimes refer to a set as being invariant, without explicitly identifying the operator under which the set is invariant.

Remark An absorbing set B as defined in Definition 2.2.1 need not be forward invariant in the sense of Definition 2.3.2. However since B is an absorbing set there exists $t^* \geq 0$ such that $S(t)B \subseteq B$ for all $t \geq t^*$. Then

$$B^* = \bigcup_{t \in [0, t^*]} S(0, t)B$$

will define a forward invariant absorbing set, and so in what follows we may assume without loss of generality that absorbing sets are forward invariant.

We now define the ω -limit set of a point and a set.

Definition 2.3.3 For any $y_0 \in U$ the ω -limit set of y_0 is defined by

$$\omega(y_0) = \bigcap_{\tau \geq 0} \overline{\bigcup_{t \geq \tau} S(t)y_0}.$$

For a bounded set $E \subseteq U$ we define the ω -limit set of E by

$$\omega(E) = \bigcap_{\tau \geq 0} \overline{\bigcup_{t \geq \tau} S(t)E}.$$

The ω -limit set $\omega(\mathbf{y}_0)$ defines the asymptotic behaviour of the trajectory with initial condition \mathbf{y}_0 . For autonomous systems $\omega(\mathbf{y}_0)$ is typically a fixed point, periodic orbit, torus, quasi-periodic orbit or strange-attractor. See [24] for a description of these objects.

Similarly $\omega(E)$ defines the asymptotic behaviour for the evolution of the set E under S . Note that

$$\bigcup_{\mathbf{x} \in E} \omega(\mathbf{x}) \subseteq \omega(E),$$

but that in general this inclusion is sharp. This is because $\omega(E)$ contains heteroclinic and homoclinic connections between the limit sets of individual trajectories originating in E .

Since we are interested in the asymptotic behaviour of dynamical systems, and this is described by the ω -limit sets, these objects will be of vital interest throughout this thesis. In the following result we state well known properties of ω -limit sets which we will need (see [28, 58, 59] for proof).

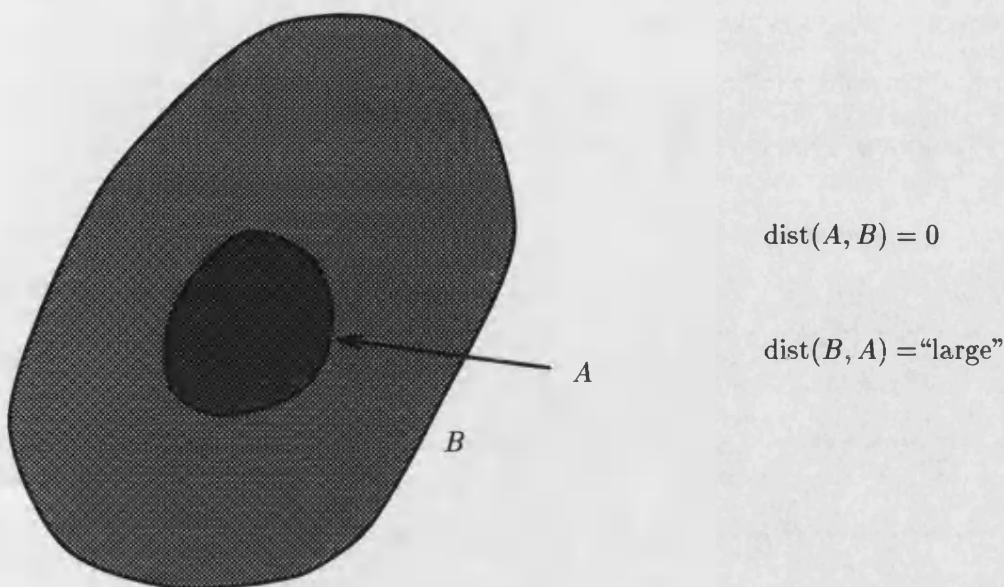
Result 2.3.4 *Suppose that (2.1.1) defines a dynamical system on $U \subseteq \mathbb{R}^m$ and that f is locally Lipschitz. If the forward orbit of \mathbf{y}_0 , $\Gamma^+(\mathbf{y}_0)$, is bounded then $\omega(\mathbf{y}_0)$ is nonempty, compact, connected and invariant under S .*

If for a bounded set $E \subseteq U$

$$\bigcup_{t \geq 0} S(t)E$$

is bounded then $\omega(E)$ is nonempty, compact and invariant under S . If E is connected, then $\omega(E)$ is also connected. ■

In Chapter 7 we will measure the distance between corresponding invariant sets of the underlying dynamical system and the discrete system generated by the discretization. The concepts of distance between sets that we will use are defined below.

Figure 2.1: The semi-distance $\text{dist}(\bullet, \bullet)$.

Definition 2.3.5 Given a set $B \subset \mathbb{R}^m$ and a point $\mathbf{x} \in \mathbb{R}^m$ we define

$$\text{dist}(\mathbf{x}, B) = \inf_{\mathbf{y} \in B} \|\mathbf{x} - \mathbf{y}\|.$$

For two sets $A, B \subset \mathbb{R}^m$ we define the semi-distance of A from B , $\text{dist}(A, B)$ by

$$\text{dist}(A, B) = \sup_{\mathbf{x} \in A} \text{dist}(\mathbf{x}, B),$$

and the Hausdorff distance between A and B , $\text{dist}_H(A, B)$, by

$$\text{dist}_H(A, B) = \max(\text{dist}(A, B), \text{dist}(B, A)).$$

Given a set A we also define the ε -neighbourhood of A by

$$\mathcal{N}(A, \varepsilon) = \{\mathbf{x} : \text{dist}(\mathbf{x}, A) < \varepsilon\}.$$

Note that in general $\text{dist}(A, B) \neq \text{dist}(B, A)$, so $\text{dist}(\bullet, \bullet)$ as defined in Definition 2.3.5 is indeed only a semi-distance. If $\text{dist}(A, B) < \varepsilon$ then $A \subseteq \mathcal{N}(A, \varepsilon)$ and if $\text{dist}(A, B) = 0$ then $A \subseteq \overline{B}$. Thus if $\text{dist}_H(A, B) = 0$, then $\overline{A} = \overline{B}$. It follows easily that the Hausdorff distance defines a metric on the set of nonempty compact subsets

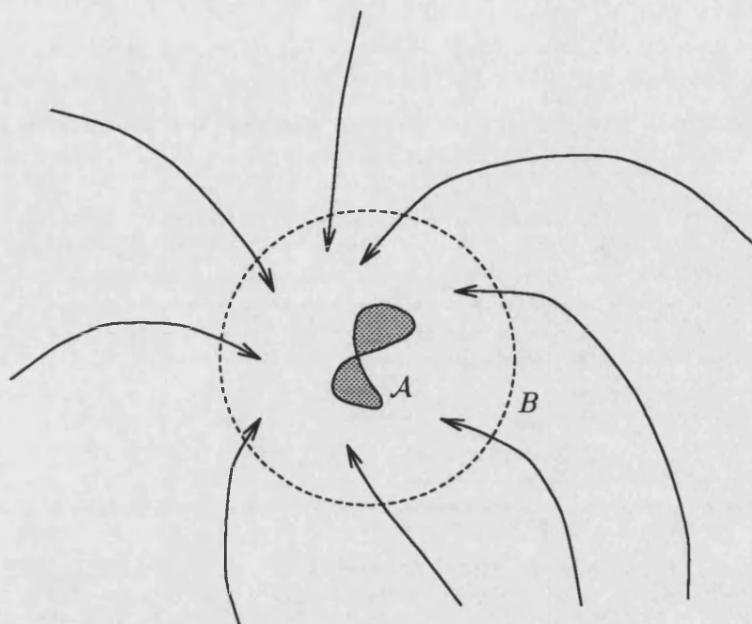


Figure 2.2: A dissipative system possesses an absorbing set B and a global attractor A .

of \mathbb{R}^m . We now define local and global attractors for a continuous dynamical system.

Definition 2.3.6 For a continuous dynamical system a set A is said to *attract a set* B under $S(t)$ if for any $\varepsilon > 0$ there exists $t^* = t^*(\varepsilon, A, B)$ such that

$$S(t)B \subseteq N(A, \varepsilon) \quad \forall t > t^*.$$

A is said to be a *local attractor* if it is a compact invariant set that attracts an open neighbourhood of itself. A is said to be a *global attractor* if it is a compact invariant set that attracts all bounded subsets of U .

We now consider the form of the global attractor A of a dissipative dynamical system. First note that since A is invariant under the flow it follows that $A \subseteq B$, where B is any absorbing set for the dynamical system. Indeed A is given by

$$A = \omega(B)$$

and hence, by Result 2.3.4, A is connected (as well as being compact and invariant).

Also note that since A attracts all bounded subsets of \mathbb{R}^m it follows that if \mathcal{I} is

a bounded invariant set then $\mathcal{I} \subseteq \mathcal{A}$, and thus the global attractor \mathcal{A} is the maximal bounded invariant set.

Although the definitions of local and global attractors are relatively straightforward, except in simple cases these objects are often hard to pin down. One approach is to recognise that the attractor is a union of invariant sets and then to identify invariant sets of the system. The simplest invariant sets of any system are its fixed points. Let \mathcal{E} be the set of equilibria of (2.1.1),

$$\mathcal{E} = \{\mathbf{x}: \mathbf{f}(\mathbf{x}) = 0\}, \quad (2.3.1)$$

and \mathcal{E}^* the set of hyperbolic equilibria,

$$\mathcal{E}^* = \{\mathbf{x} \in \mathcal{E}: \mathbf{x} \text{ is hyperbolic}\}. \quad (2.3.2)$$

Definition 2.3.7 If \mathbf{x}^* is a hyperbolic fixed point of (2.1.1) then the *unstable manifold* of \mathbf{x}^* is defined by

$$W(\mathbf{x}^*) = \{\mathbf{y}: S(t)\mathbf{y} \text{ exists } \forall t \leq 0 \text{ and } S(t)\mathbf{y} \rightarrow \mathbf{x}^* \text{ as } t \rightarrow -\infty\}.$$

For some $\delta > 0$ we define the *local unstable manifold* of \mathbf{x}^* by

$$W^\delta(\mathbf{x}^*) = \{\mathbf{y} \in W(\mathbf{x}^*): S(t)\mathbf{y} \in \overline{B}(\mathbf{x}^*, \delta) \forall t \leq 0\}.$$

Note that for a dissipative system $W(\mathbf{x}^*)$ is necessarily bounded and it follows that $W(\mathbf{x}^*)$ is invariant and hence that $W(\mathbf{x}^*) \subseteq \mathcal{A}$. Indeed let

$$W(\mathcal{E}^*) = \bigcup_{\mathbf{x}^* \in \mathcal{E}^*} W(\mathbf{x}^*) \quad (2.3.3)$$

then the following theorem shows that $\overline{W(\mathcal{E}^*)}$ is invariant and hence that $\overline{W(\mathcal{E}^*)} \subseteq \mathcal{A}$.

Theorem 2.3.8 Let $W(\mathcal{E}^*)$ be defined by (2.3.3) then if $W(\mathcal{E}^*)$ is bounded (and in particular, if the system is dissipative) then $\overline{W(\mathcal{E}^*)}$ is compact and invariant under $S(\bullet)$.

Proof. Let $B = W(\mathcal{E}^*)$. Then

$$\begin{aligned}\omega(B) &= \bigcap_{\tau \geq 0} \overline{\bigcup_{t \geq \tau} S(t)B} \\ &= \overline{B}\end{aligned}$$

since B is invariant under $S(\bullet)$. Now the result follows from Result 2.3.4. ■

In Section 2.4 we will consider gradient dynamical systems. For a dissipative gradient system with hyperbolic fixed points the global attractor is known (see Hale [28]) to be of the form $\mathcal{A} = \overline{W(\mathcal{E}^*)}$. In Example 2.3.9 below the global attractor also has this form, although the system is not in gradient form. However, in general, for nongradient systems $\overline{W(\mathcal{E}^*)} \subset \mathcal{A}$ with the inclusion being strict. In the case where $\overline{W(\mathcal{E}^*)} \subset \mathcal{A}$ with strict inclusion we could continue to build up the attractor by considering more complicated invariant sets such as periodic orbits and tori together with their invariant manifolds, but we will not pursue this, due to lack of space and time.

Finally in this section we present an illustrative example.

Example 2.3.9 Recall from Example 2.2.7 that (2.2.14) defines a dissipative system. This system has the exact solution

$$\begin{aligned}r(t) &= \frac{1}{\sqrt{1 + \frac{1-r_0^2}{r_0^2} e^{-2t}}} \\ \theta(t) &= \theta_0 + t\end{aligned}$$

for initial condition (r_0, θ_0) if $r_0 \neq 0$ and $r(t) = 0 \forall t$ if $r_0 = 0$. Note that if $r_0 = 1$ then $r(t) = 1 \forall t \in \mathbb{R}$ and the unit circle is an invariant set. Indeed if $r_0 \neq 0$ then $r(t) \rightarrow 1$ as $t \rightarrow \infty$, and it follows that the unit circle attracts all open closed bounded sets that do not contain the origin, and hence is a local attractor. Since the unit circle does not attract any set that contains the origin, it is not the global attractor of this system. It is easy to show that the closed unit disc is invariant under the evolution of (2.2.14) and attracts all bounded subsets of \mathbb{R}^m , and hence is the global attractor of this system.

The system has one hyperbolic fixed point at the origin. The open unit disc is the unstable manifold of this fixed point, and hence $\mathcal{A} = \overline{W(\mathcal{E}^*)}$, that is the global attractor of the system is equal to the closure of the unstable manifold of the hyperbolic fixed point. See Figure 2.3. □

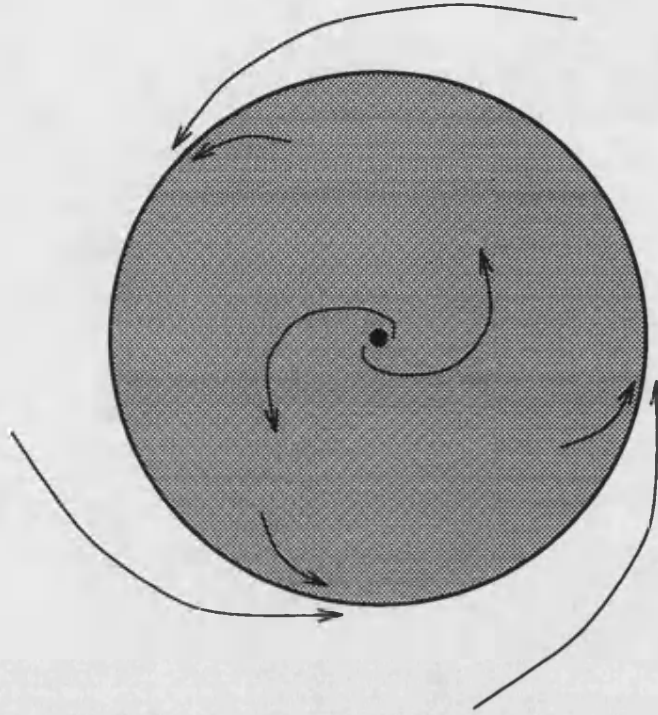


Figure 2.3: The dynamics of (2.2.14). The unit circle is a local attractor and the closed unit disc is the global attractor.

2.4 Gradient Dynamical Systems

In Section 2.1 we noted that if f satisfies a one-sided Lipschitz condition with $c < 0$ then all trajectories of the system are asymptotic to a unique fixed point as $t \rightarrow \infty$. A natural generalization of these problems is to consider systems which possess multiple fixed points, and for which every trajectory is asymptotic to some fixed point. Gradient systems, as defined below, are dynamical systems which have this property, and we will consider their numerical approximation in Chapter 5.

Definition 2.4.1 If (2.1.1) defines a dynamical system on $U \subseteq \mathbb{R}^m$ then (2.1.1) is said to define a *gradient system* on U if $\exists F: U \rightarrow \mathbb{R}$ satisfying

- (i) $F(\bullet)$ is bounded below on U ,
- (ii) $F(y) \rightarrow \infty$ as $\|y\| \rightarrow \infty$,
- (iii) $F(S(t)y_0)$ is non-increasing in t for a solution of (2.1.1), and,
- (iv) if $F(S(t)y_0) = F(y_0)$ for $t > 0$ then $y(0)$ is an equilibrium point.

F is called a *Lyapunov functional*.

Note that if $U = \mathbb{R}^m$, \mathbf{f} is locally Lipschitz, and (ii) and (iii) hold then all trajectories are bounded and it follows from the remarks in Section 2.1 that (2.1.1) does define a dynamical system, hence if (i) and (iv) also hold then (2.1.1) defines a gradient system.

Let \mathcal{E} be the set of equilibria of (2.1.1), as defined in (2.3.1). The following result shows that every trajectory must converge to a fixed point of the system. Hirsch and Smale [34] prove a similar result, but under the slightly stronger assumption that \mathbf{f} is C^1 . The proof is very similar to but simpler than that of Theorem 2.5.4 below, and is thus omitted.

Result 2.4.2 *If (2.1.1) defines a gradient system and \mathbf{f} is locally Lipschitz then $\omega(\mathbf{y}_0) \subseteq \mathcal{E}$. If, furthermore, the zeros of \mathbf{f} are isolated then $\omega(\mathbf{y}_0) = \mathbf{x}$ for some $\mathbf{x} \in \mathcal{E}$.*

■

Note that if \mathbf{f} satisfies

$$\mathbf{f}(\mathbf{y}) = -\nabla F(\mathbf{y}) \quad (2.4.1)$$

for some $F: U \rightarrow \mathbb{R}$ then

$$\begin{aligned} \frac{d}{dt}F(\mathbf{y}(t)) &= \langle \mathbf{f}(\mathbf{y}), \nabla F(\mathbf{y}(t)) \rangle \\ &= -\|\mathbf{f}(\mathbf{y})\|^2 \end{aligned}$$

and thus (iii) and (iv) of Definition 2.4.1 follow automatically. Hence if F satisfies (i) and (ii) then (2.1.1, 2.4.1) defines a gradient system.

Definition 2.4.1 (iii) and (iv) imply that any solution trajectory of a gradient system must travel down hill on a contour map of F . The additional condition (2.4.1) implies that trajectories not only travel down hill, but must also follow the path of steepest descent. Condition (2.4.1) is therefore a natural one for gradient systems to satisfy, and systems which arise in applications almost always satisfy this or a closely related condition; indeed some authors even include (2.4.1) as part of their definition of a gradient system (see [34] for example). Thus when we come to consider the numerical approximation of gradient dynamical systems there will be little loss in generality in assuming that \mathbf{f} satisfies (2.4.1) and so we will make this assumption.

We now prove two lemmas which will be useful when we consider the numerical solution of (2.1.1, 2.4.1) in Chapter 5. We begin by deriving an upper bound for $F(\mathbf{u}) - F(\mathbf{v})$ when \mathbf{f} satisfies a one-sided or global Lipschitz condition.

Lemma 2.4.3 *If f satisfies a one-sided Lipschitz condition (2.1.6) on a convex set $B \subseteq U$ then the gradient system (2.1.1, 2.4.1) satisfies*

$$F(u) - F(v) \leq \langle f(u), v - u \rangle + c\|v - u\|^2 \quad (2.4.2)$$

for all $u, v \in B$.

Proof. Let $G(x): [0, 1] \rightarrow \mathbb{R}$ be defined by

$$G(x) = F(v + x[u - v]).$$

Then we have

$$\begin{aligned} G'(x) &= \langle \nabla F(v + x[u - v]), u - v \rangle \\ &= \langle f(v + x[u - v]), v - u \rangle. \end{aligned}$$

Now by the mean value theorem $G(1) - G(0) = G'(x)$ for some $x \in (0, 1)$. Hence writing $\xi = v + x[u - v]$ implies that

$$F(u) - F(v) = \langle f(\xi), v - u \rangle \quad (2.4.3)$$

and since

$$\frac{v - u}{\|v - u\|} = \frac{\xi - u}{\|\xi - u\|}$$

it follows that

$$F(u) - F(v) = \frac{\|v - u\|}{\|\xi - u\|} \langle f(\xi), \xi - u \rangle$$

Now substituting ξ for v in (2.1.6) implies that

$$\langle f(\xi), \xi - u \rangle \leq \langle f(u), \xi - u \rangle + c\|u - \xi\|^2$$

and hence

$$\begin{aligned} F(u) - F(v) &\leq \frac{\|v - u\|}{\|\xi - u\|} \langle f(u), \xi - u \rangle + c\|\xi - u\| \cdot \|v - u\| \\ &\leq \frac{\|v - u\|}{\|\xi - u\|} \langle f(u), \xi - u \rangle + c\|v - u\|^2 \\ &= \langle f(u), v - u \rangle + c\|v - u\|^2 \end{aligned}$$

as required. ■

Note that we cannot obtain greater generality by considering gradient systems under the assumption that (2.4.2) holds, instead of assuming a one-sided Lipschitz condition, because (2.4.2) is essentially equivalent to the one-sided Lipschitz condition (2.1.6). To see this notice that exchanging \mathbf{u} and \mathbf{v} in (2.4.2) implies that

$$F(\mathbf{v}) - F(\mathbf{u}) \leq \langle \mathbf{f}(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle + c\|\mathbf{v} - \mathbf{u}\|^2$$

and adding this with (2.4.2) we obtain

$$0 \leq \langle \mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{v}), \mathbf{v} - \mathbf{u} \rangle + 2c\|\mathbf{u} - \mathbf{v}\|^2,$$

or on rearranging

$$\langle \mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle \leq 2c\|\mathbf{u} - \mathbf{v}\|^2.$$

Thus \mathbf{f} satisfies (2.4.2) if and only if \mathbf{f} satisfies a one-sided Lipschitz condition.

If \mathbf{f} is globally Lipschitz then in addition to (2.4.2) we also obtain inequality (2.4.4) below.

Lemma 2.4.4 *If \mathbf{f} is Lipschitz with Lipschitz constant L on a convex set $B \subseteq U$ then the gradient system (2.1.1, 2.4.1) satisfies*

$$F(\mathbf{u}) - F(\mathbf{v}) \leq \langle \mathbf{f}(\mathbf{v}), \mathbf{v} - \mathbf{u} \rangle + L\|\mathbf{v} - \mathbf{u}\|^2 \quad (2.4.4)$$

for all $\mathbf{u}, \mathbf{v} \in B$.

Proof. Derive (2.4.3) as in the proof of Lemma 2.4.3. Then since

$$\frac{\mathbf{v} - \mathbf{u}}{\|\mathbf{v} - \mathbf{u}\|} = \frac{\mathbf{v} - \boldsymbol{\xi}}{\|\mathbf{v} - \boldsymbol{\xi}\|}$$

it follows that

$$F(\mathbf{u}) - F(\mathbf{v}) = \frac{\|\mathbf{v} - \mathbf{u}\|}{\|\mathbf{v} - \boldsymbol{\xi}\|} \langle \mathbf{f}(\boldsymbol{\xi}), \mathbf{v} - \boldsymbol{\xi} \rangle. \quad (2.4.5)$$

Now since \mathbf{f} is globally Lipschitz and $\boldsymbol{\xi} \in B$ by convexity, use of the Cauchy-Schwarz inequality gives

$$\langle \mathbf{f}(\boldsymbol{\xi}) - \mathbf{f}(\mathbf{v}), \mathbf{v} - \boldsymbol{\xi} \rangle \leq \|\mathbf{f}(\boldsymbol{\xi}) - \mathbf{f}(\mathbf{v})\| \cdot \|\mathbf{v} - \boldsymbol{\xi}\|$$

$$\leq L\|\xi - v\|^2$$

and hence

$$\langle f(\xi), v - \xi \rangle \leq \langle f(v), v - \xi \rangle + L\|v - \xi\|^2.$$

Now applying this to (2.4.5) implies

$$\begin{aligned} F(u) - F(v) &\leq \frac{\|v - u\|}{\|v - \xi\|} \langle f(v), v - \xi \rangle + L\|v - \xi\| \cdot \|v - u\| \\ &\leq \frac{\|v - u\|}{\|v - \xi\|} \langle f(v), v - \xi \rangle + L\|v - u\|^2 \\ &= \langle f(v), v - u \rangle + L\|v - u\|^2 \end{aligned}$$

as required. ■

2.5 Discrete Dynamical Systems

Next we define a dynamical systems for mappings. We will consider a family of maps, Φ_h , parameterized by h . When we consider the numerical approximation of (2.1.1) by a Runge-Kutta method (3.2.1–2) the parameter h will be the step-size of the method.

Definition 2.5.1 Suppose $\Phi_h: U \rightarrow U$ where $U \subseteq \mathbb{R}^m$ then for any $y_0 \in U$ the sequence $\{y_n\}_{n=0}^\infty$ is uniquely defined in U by

$$y_{n+1} = \Phi_h(y_n), \quad (2.5.1)$$

and Φ_h is said to define a *discrete dynamical system* on U . The evolution operator $S_h: U \rightarrow U$ is defined to be the operator such that $y_{n+1} = S_h y_n$, and hence $S_h \equiv \Phi_h$. We denote the n -fold composition of S_h by S_h^n , so that $y_n = S_h^n y_0$, and n plays the role of a discrete time variable.

We can define continuity, dissipativity, orbits, invariance, ω -limit sets, unstable manifolds and attractors for discrete dynamical systems in the natural way by replacing $S(t)$ with S_h^n in the relevant definitions with n playing the role of t for integer n . Since the definitions for discrete dynamical systems are completely analogous to those for dynamical systems we will not restate them. To avoid confusion between ω -limit sets, attractors, and other invariant sets for discrete dynamical systems with the corre-

sponding sets for dynamical systems we will always denote the objects associated with the discrete dynamical system with a subscript h . Hence $\omega_h(\mathbf{y}_0)$ denotes the ω -limit set of \mathbf{y}_0 under the evolution of (2.5.1) and \mathcal{A}_h denotes an attractor of (2.5.1).

The following result is analogous to Result 2.3.4 and will be needed in the next section. See [28], for example, for proof.

Theorem 2.5.2 *Suppose that (2.5.1) defines a discrete dynamical system on $U \subseteq \mathbb{R}^m$ and that Φ_h is locally Lipschitz. If the forward orbit of \mathbf{y}_0 , $\overline{\bigcup_{n \geq 0} S_h^n \mathbf{y}_0}$ is bounded then $\omega_h(\mathbf{y}_0)$ is nonempty, compact and invariant under S_h . ■*

2.5.1 Discrete Gradient Systems

The definition of a discrete dynamical system is analogous to Definition 2.4.1.

Definition 2.5.3 If (2.5.1) defines a discrete dynamical system on $U \subseteq \mathbb{R}^m$ then (2.5.1) is said to define a *discrete gradient system* if $\exists F_h: U \rightarrow \mathbb{R}$ satisfying

- (i) $F_h(\bullet)$ is bounded below on U ,
- (ii) $F_h(\mathbf{y}) \rightarrow \infty$ as $\|\mathbf{y}\| \rightarrow \infty$,
- (iii) $F_h(S_h^n \mathbf{y}_0)$ is nonincreasing in n for a solution of (2.1.1), and,
- (iv) if $F_h(S_h \mathbf{y}) = F_h(\mathbf{y})$ then \mathbf{y} is an equilibrium point of (2.5.1).

F_h is called a *Lyapunov functional*.

The discrete gradient systems that we will consider arise as numerical approximations to gradient systems for which \mathbf{f} is locally Lipschitz and this will imply that Φ_h is locally Lipschitz. For such systems we can prove an analogous result to Result 2.4.2 as follows; see [16] and [20] for related results. Let $\mathcal{E}_h = \{\mathbf{y}: \Phi_h(\mathbf{y}) = \mathbf{y}\}$, the set of equilibria of (2.5.1).

Theorem 2.5.4 *If (2.5.1) defines a discrete gradient system with Φ_h locally Lipschitz on U , then $\omega_h(\mathbf{y}_0) \subseteq \mathcal{E}_h$. If, furthermore, the fixed points of Φ_h are isolated then $\omega_h(\mathbf{y}_0) = \mathbf{x}$ for some $\mathbf{x} \in \mathcal{E}_h$.*

Proof. Since, by property (iii), $F_h(S_h^n \mathbf{y}_0) \leq F_h(\mathbf{y}_0)$ and property (ii) holds, it follows that $\overline{\bigcup_{n \geq 0} S_h^n \mathbf{y}_0}$ is compact, and hence since Φ_h is locally Lipschitz Theorem 2.5.2 implies that $\omega_h(\mathbf{y}_0)$ is non-empty, compact and invariant.

If $\mathbf{x}_1, \mathbf{x}_2 \in \omega(\mathbf{y}_0)$ then it is clear that $F_h(\mathbf{x}_1) = F_h(\mathbf{x}_2)$ for otherwise we obtain a contradiction to property (iii). Since $\omega_h(\mathbf{y}_0)$ is invariant it follows that $F_h(S_h^n \mathbf{x}) = F_h(\mathbf{x})$ for any $\mathbf{x} \in \omega_h(\mathbf{y}_0)$. Thus $\mathbf{x} \in \mathcal{E}_h$ by property (iv).

Now assume that the fixed points of Φ_h are isolated. Since $\omega_h(y_0)$ is compact it follows that $\omega_h(y_0)$ contains a finite number of equilibria, say $x_j, j = 1, \dots, N$. Define

$$\Delta = \min_{i \neq j} \text{dist}(x_i, x_j),$$

and let $L \geq 1$ be the Lipschitz constant for Φ_h on $B(x_1, \Delta)$. Now let $\delta = \Delta/2L$ and assume for the purposes of contradiction that $N \geq 2$. Since $x_1 \in \omega_h(y_0)$ there exist infinitely many n such that $S_h^n y_0 \in B(x_1, \delta)$. But since x_1 is not the unique point of $\omega_h(y_0)$ there exist infinitely many n such that $S_h^n y_0 \notin B(x_1, \delta)$. Hence we may construct an infinite sequence of integers $n_i \rightarrow \infty$ such that $S_h^{n_i} y_0 \in B(x_1, \delta)$ and $S_h^{n_i+1} y_0 \notin B(x_1, \delta)$. Notice that

$$\begin{aligned} \|y_{n_i+1} - x_1\| &= \|\Phi_h(y_{n_i}) - \Phi_h(x_1)\| \\ &\leq L\|y_{n_i} - x_1\| \\ &< L\delta \\ &= \Delta/2. \end{aligned}$$

Hence $y_{n_i+1} \notin B(x_j, \Delta/2)$ for any i or j . But $\{y_{n_i+1}\}_{i=0}^\infty$ is bounded, and hence must have a limit point; but by construction such a limit point cannot be contained in \mathcal{E}_h and hence we have obtained a contradiction. This completes the proof. ■

2.5.2 Implicit Maps

Suppose that y_{n+1} is not given by an explicit mapping of the form (2.5.1), but instead y_{n+1} is obtained from y_n via an implicit mapping of the form

$$\Upsilon_h(y_n, y_{n+1}) = 0. \quad (2.5.2)$$

Implicit numerical methods applied to (2.1.1) define such mappings. For a given y_n there may be none, one or many solutions to (2.5.2), and hence in general these maps do not define dynamical systems, and will not have well defined evolution operators S_h . However we can define a generalized evolution operator for (2.5.2) as follows.

Definition 2.5.5 We define the *generalized evolution operator* G_h for (2.5.2) by

$$\begin{aligned} G_h(\mathbf{u}) &= \{\mathbf{v}: \Upsilon_h(\mathbf{u}, \mathbf{v}) = 0\}, \\ G_h(E) &= \bigcup_{\mathbf{u} \in E} G_h^1(\mathbf{u}), \end{aligned}$$

and the n -fold composition of G_h , denoted by G_h^n , is given by

$$\begin{aligned} G_h^n(\mathbf{u}) &= G_h(G_h^{n-1}(\mathbf{u})), \\ G_h^n(E) &= \bigcup_{\mathbf{u} \in E} G_h^n(\mathbf{u}). \end{aligned}$$

The generalized evolution operator G_h is the natural extension of the evolution operator S_h to multi-valued maps. If (2.5.2) is uniquely soluble for all $\mathbf{y} \in \mathbb{R}^m$ then (2.5.2) defines a dynamical system on \mathbb{R}^m and S_h and G_h agree. In general however, the evolution operator $G_h(\mathbf{y}_n)$ returns all the possible values of \mathbf{y}_{n+1} for the implicit map, and should be thought of as a set-valued function on subsets of \mathbb{R}^m rather than a map from \mathbb{R}^m to itself. Note that if (2.5.2) is insoluble for some initial condition \mathbf{y}_n then $G_h(\mathbf{y}_n) = \emptyset$, and we also define $G_h(\emptyset) = \emptyset$.

It should be noted that our definition of the generalized evolution operator is analogous to the usual definition of negative orbits for discrete dynamical systems: since the map defining a discrete dynamical system need not be one-to-one, negative orbits need not be unique, and it is usual, see for example Hale [28], to define the negative orbit of a point to be the union of all possible such orbits.

The generalized evolution map allows us to extend the concept of dissipativity to cover multi-valued maps, in a natural way, by replacing S_h^n by G_h^n in the definition of dissipativity. We will use this generalized concept of dissipativity in Chapter 6.

2.6 Fixed Point Theorems

We will later require Brouwer's Fixed Point Theorem and state it here without proof.

Result 2.6.1 (Brouwer's Fixed Point Theorem) *If $f: B \rightarrow \mathbb{R}^m$ is continuous, B is a nonempty compact convex subset of \mathbb{R}^m and $f(B) \subseteq B$ then there exists $\mathbf{x} \in B$ such that $f(\mathbf{x}) = \mathbf{x}$. ■*

It follows trivially from Brouwer's Fixed Point Theorem that if Φ_h is continuous then any set which is compact, convex and forward invariant under the evolution of the discrete dynamical system (2.5.1) contains a fixed point. We can also derive the same result for (2.1.1) from Brouwer's Fixed Point Theorem.

Result 2.6.2 *If f is Lipschitz on B and B is a compact convex forward invariant set under the evolution of the dynamical system (2.1.1) then B contains a fixed point.*

Proof. For $h > 0$ let $S_h = S(h)|_B$ then it follows from the Lipschitz continuity of f that S_h is the evolution operator of a continuous discrete dynamical system, and B is forward invariant under the evolution of this system. Therefore by Result 2.6.1 there exists $x(h) \in B$ such that $S_h x(h) = x(h)$. This is true for all $h > 0$ and hence we can choose a sequence $\{h_n\}_{n=0}^\infty$ with $h_n \rightarrow 0$ as $n \rightarrow \infty$ such that there exists a sequence $\{x_n\}_{n=0}^\infty$ with $S_{h_n} x_n = x_n$ and $x_n \rightarrow x^*$ as $n \rightarrow \infty$ for some $x^* \in B$. We claim that x^* is a fixed point of (2.1.1). Suppose not, then there exists $t^* > 0$ and $y \in B$ such that $S(t^*)x^* = y$ with $\|x^* - y\| \geq \varepsilon$ for some $\varepsilon > 0$. Now by continuity with respect to initial data there exists $t_1 < t^* < t_2$ and a neighbourhood N of x^* such that if $x \in N$ and $t \in (t_1, t_2)$ then $\|S(t)x - y\| \leq \varepsilon/2$. But this implies that if $x \in N$ is on a periodic orbit of (2.1.1) then the period is greater than $t_2 - t_1$, which provides the required contradiction, since $\{x_n\}_{n=0}^\infty$ is a sequence of periodic points of (2.1.1) converging to x^* with period tending to zero as $n \rightarrow \infty$. ■

Chapter 3

Numerical Methods

3.1 Introduction

In this chapter we will define the numerical methods that we shall use. The simplest numerical method for the solution of (2.1.1) is the forward Euler method, defined by

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}(\mathbf{y}_n), \quad (3.1.1)$$

where h is the (constant) step-size, $t_n = nh$ and \mathbf{y}_n is an approximation to $\mathbf{y}(t_n)$.

There are two natural ways to generalize the forward Euler method to obtain more sophisticated methods. One approach is to store and use the values of \mathbf{y}_n and $\mathbf{f}(\mathbf{y}_n)$ from previous steps. This results in the so-called *multistep methods* described in Section 3.3 below. Another approach is not to use previous values, but to evaluate \mathbf{f} at intermediate stages between \mathbf{y}_n and \mathbf{y}_{n+1} , resulting in one-step multistage methods. The one-step methods that we will consider throughout are the Runge-Kutta methods.

Recently there has been some interest in the literature in multistage multistep methods; see, for example, Butcher [9] where such methods are referred to as general linear methods. If derivatives of \mathbf{f} are also used then we obtain multi-derivative multistage multistep methods. However methods which use derivatives of \mathbf{f} are not popular because of the cost of differentiation, although with the increasing proliferation of symbolic mathematics packages this could change.

Although we will not consider variable time-stepping strategies, it should be noted that a serious disadvantage of multistep and multistage multistep methods is that in variable time-stepping implementations when the step-size is changed previous values

of \mathbf{y}_n and \mathbf{f} will not be directly available and, in general, will have to be approximated using an interpolation routine; thus increasing both the error and the cost of the method.

We will mainly concentrate on the case where (2.1.1) is solved numerically by Runge-Kutta methods, but we will also consider the solution of (2.1.1) by linear multistep methods. Multistage multistep methods have yet to be included in popular integration packages, and we will not consider them further.

3.2 Runge-Kutta Methods

A general s -stage fixed time-stepping Runge-Kutta method for the solution of (2.1.1) may be written as:

$$\mathbf{Y}_i = \mathbf{y}_n + h \sum_{j=1}^s a_{ij} \mathbf{f}(\mathbf{Y}_j), \quad i = 1, \dots, s \quad (3.2.1)$$

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h \sum_{i=1}^s b_i \mathbf{f}(\mathbf{Y}_i). \quad (3.2.2)$$

Here \mathbf{y}_n approximates the exact solution $\mathbf{y}(t_n)$ at $t_n = nh$, where $h > 0$ is the fixed step-size. Runge-Kutta methods are often represented using the Butcher tableau

$$\begin{array}{c|c} \mathbf{c} & \mathbf{A} \\ \hline & \mathbf{b}^T \end{array} \equiv \begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \dots & a_{1s} \\ c_2 & a_{21} & a_{22} & \dots & a_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & a_{s2} & \dots & a_{ss} \\ \hline & b_1 & b_2 & \dots & b_s \end{array} \quad (3.2.3)$$

where

$$c_i = \sum_{j=1}^s a_{ij}. \quad (3.2.4)$$

We will always assume that the method is consistent, which implies that

$$\sum_{i=1}^s b_i = 1. \quad (3.2.5)$$

We will also use the notation

$$a = \max_i \sum_{j=1}^{i-1} |a_{ij}| + \max_i \sum_{j=i}^s |a_{ij}|, \quad (3.2.6)$$

$$\mathbf{A} = \max_i \sum_{j=1}^s |a_{ij}| = \|\mathbf{A}\|_\infty, \quad (3.2.7)$$

and

$$\mathbf{B} = \sum_{i=1}^s |b_i| \geq 1. \quad (3.2.8)$$

Notice that

$$\mathbf{A} \leq a \leq 2\mathbf{A} \quad (3.2.9)$$

which implies that

$$\frac{1}{a} \geq \frac{1}{\mathbf{A}(1 + \mathbf{B})}. \quad (3.2.10)$$

The method (3.2.1–2) is said to be *explicit* if

$$a_{i,j} = 0 \quad \forall 1 \leq i \leq j \leq s$$

and *implicit* otherwise. For an implicit method we will always assume that the defining equations (3.2.1) are solved exactly.

We will require two $s \times s$ matrices B and M associated with the Runge-Kutta method (3.2.1–2), and given by

$$B = \text{diag}(b_1, b_2, \dots, b_s), \quad (3.2.11)$$

$$M = BA + A^T B - \mathbf{b}\mathbf{b}^T. \quad (3.2.12)$$

We will denote the ij -th entry of M by m_{ij} . Hence

$$m_{ij} = b_i a_{ij} + b_j a_{ji} - b_i b_j. \quad (3.2.13)$$

Using these matrices, and following Burrage and Butcher [6], we make the following definition.

Definition 3.2.1 A Runge-Kutta method is said to be *algebraically stable* if the two matrices B and M defined by (3.2.11–12) are both positive semi-definite.

This concept will play an important role in this thesis, as will become apparent.

In Chapter 6 and in Proposition 3.2.5 we will also require the $s \times s$ matrix E

associated with (3.2.1–2) and defined by $E = \{e_{ij}\}$ where

$$e_{ij} = b_j - a_{ij}. \quad (3.2.14)$$

3.2.1 Reducibility

It is possible for a Runge-Kutta method to have redundant stages which do not affect the solution, and such methods are said to be reducible. There are two different concepts of reducibility for Runge-Kutta methods. We will only use the concept of DJ-reducibility, first defined by Dahlquist and Jeltsch, (see [13] or [27]).

Definition 3.2.2 A Runge-Kutta method is said to be *DJ-reducible*, if for some non-empty index set $T \subset \{1, \dots, s\}$,

$$b_j = 0 \quad \text{for } j \in T \quad \text{and} \quad a_{ij} = 0 \quad \text{for } i \notin T, j \in T,$$

and is said to be *DJ-irreducible* otherwise.

For a DJ-reducible method, the stages for which $j \in T$ do not affect the solution, and so we can define an essentially equivalent DJ-irreducible method with fewer stages by deleting the redundant stages of the DJ-reducible method. For this reason DJ-reducible methods are not used in practice, and there is no loss of generality in only considering DJ-irreducible methods.

It is well known that for a DJ-irreducible algebraically stable Runge-Kutta method $b_i > 0$ for all $i = 1, \dots, s$. Hence an algebraically stable Runge-Kutta method with $b_i = 0$ for some i must be DJ-reducible.

We also note here that an \tilde{s} -stage DJ-irreducible method formed by deleting the redundant stages of an s -stage DJ-reducible algebraically stable method will also be algebraically stable. To see this note that

$$\mathbf{x}^T \widetilde{M} \mathbf{x} = \tilde{\mathbf{x}}^T M \tilde{\mathbf{x}}$$

where \widetilde{M} is the M-matrix (3.3.1) of the reduced method, M is the M-matrix of the original method, \mathbf{x} is an arbitrary vector of dimension \tilde{s} and $\tilde{\mathbf{x}}$ is a vector of dimension s formed by inserting $s - \tilde{s}$ zeros into \mathbf{x} in positions corresponding to the redundant stages of the original method. Since the original method is algebraically stable, it now

follows that $\mathbf{x}^T \widetilde{\mathbf{M}} \mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^s$, and the reduced method is also algebraically stable.

The other concept of reducibility for Runge-Kutta methods is S-reducibility. We will mention S-reducibility briefly in Section 3.4, but since we will not actually use this concept, we will not define it explicitly. See [13] or [27] for a formal definition of S-reducibility. We note here that all non-confluent ($c_i \neq c_j$ for $i \neq j$) methods are S-irreducible (see [13]), and that this includes all the commonly used methods.

3.2.2 Properties of Runge-Kutta Solutions

In this section we will present results which specify properties of solutions to the Runge-Kutta defining equations (3.2.1–2), which will be useful in this and subsequent chapters. Notice that if $\mathbf{A} = 0$ then (3.2.7) implies that $a_{ij} = 0 \forall i, j$ and hence $\mathbf{f}(\mathbf{Y}_i) = \mathbf{f}(\mathbf{y}_n)$ for all i and

$$\|\mathbf{f}(\mathbf{y}_n) - \sum_{i=1}^s b_i \mathbf{f}(\mathbf{Y}_i)\| \equiv 0. \quad (3.2.15)$$

Thus the solution sequence of the method is equivalent to that of the forward Euler method. The following lemma establishes a bound on $\|\mathbf{f}(\mathbf{y}_n) - \mathbf{f}(\mathbf{Y}_i)\|$ when $\mathbf{A} \neq 0$.

Lemma 3.2.3 *If \mathbf{f} is Lipschitz on $B \subseteq \mathbb{R}^m$ with Lipschitz constant L , $\mathbf{y}_n \in B$, $\mathbf{A} > 0$ and $h < 1/L\mathbf{A}$ then any solution of the Runge-Kutta defining equations (3.2.1–2) which satisfies $\mathbf{Y}_i \in B$ for all i also satisfies*

$$\|\mathbf{f}(\mathbf{y}_n) - \mathbf{f}(\mathbf{Y}_i)\| < \frac{L\mathbf{A}h}{1 - L\mathbf{A}h} \|\mathbf{f}(\mathbf{y}_n)\| \quad \forall i = 1, \dots, s \quad (3.2.16)$$

where \mathbf{A} and \mathbf{B} are defined by (3.2.7) and (3.2.8).

Proof. Consider the equations (3.2.1–2). Let

$$M = \max_j \|\mathbf{f}(\mathbf{Y}_j)\|$$

then

$$\begin{aligned} \|\mathbf{f}(\mathbf{y}_n) - \mathbf{f}(\mathbf{Y}_i)\| &= \|\mathbf{f}(\mathbf{y}_n) - \mathbf{f}(\mathbf{y}_n + h \sum_{j=1}^s a_{i,j} \mathbf{f}(\mathbf{Y}_j))\| \\ &\leq Lh \left\| \sum_{j=1}^s a_{i,j} \mathbf{f}(\mathbf{Y}_j) \right\| \\ &\leq Lh\mathbf{A}M \end{aligned} \quad (3.2.17)$$

Hence

$$\|f(Y_i)\| \leq LhAM + \|f(y_n)\|$$

and in particular

$$\begin{aligned} M &\leq LhAM + \|f(y_n)\| \\ (1 - LAh)M &\leq \|f(y_n)\| \end{aligned}$$

and the result follows from (3.2.17). ■

We now derive a bound on $\|f(y_n) - \sum_{i=1}^s b_i f(Y_i)\|$ when $A \neq 0$.

Lemma 3.2.4 *If f is Lipschitz on $B \subseteq \mathbb{R}^m$ with Lipschitz constant L , $y_n \in B$ and*

$$h < \frac{1}{LA(1+B)}, \quad (3.2.18)$$

where $A > 0$, then any solution of the Runge-Kutta defining equations (3.2.1–2) which satisfies $Y_i \in B$ for all i also satisfies

$$\|f(y_n) - \sum_{i=1}^s b_i f(Y_i)\| < \frac{LAB}{1 - LAh(1+B)} \|y_{n+1} - y_n\| \quad (3.2.19)$$

where A and B are defined by (3.2.7) and (3.2.8).

Proof. Recalling (3.2.5) and applying Lemma 3.2.3 we obtain

$$\begin{aligned} \|f(y_n) - \sum_{i=1}^s b_i f(Y_i)\| &= \left\| \sum_{i=1}^s b_i (f(y_n) - f(Y_i)) \right\| \\ &\leq B \max_i \|f(y_n) - f(Y_i)\| \\ &\leq \frac{LhAB}{1 - LhA} \|f(y_n)\|. \end{aligned} \quad (3.2.20)$$

Note also that

$$\begin{aligned} \frac{1}{h} \|y_{n+1} - y_n\| &= \left\| \sum_{i=1}^s b_i f(Y_i) \right\| \\ &\geq \|f(y_n)\| - \left\| f(y_n) - \sum_{i=1}^s b_i f(Y_i) \right\| \end{aligned}$$

and hence by (3.2.20)

$$\frac{1}{h} \|\mathbf{y}_{n+1} - \mathbf{y}_n\| \geq \frac{1 - LhA(1 + B)}{1 - LhA} \|f(\mathbf{y}_n)\|. \quad (3.2.21)$$

Now the result follows on combining (3.2.20) and (3.2.21). ■

Finally in this section we establish two identities that any solution of the Runge-Kutta defining equations (3.2.1–2) must satisfy.

Proposition 3.2.5 *Any solution of the Runge-Kutta defining equations applied to (2.1.1) satisfies*

$$\|\mathbf{y}_{n+1}\|^2 = \|\mathbf{y}_n\|^2 + 2h \sum_{i=1}^s b_i \langle \mathbf{Y}_i, f(\mathbf{Y}_i) \rangle - h^2 \sum_{i,j=1}^s m_{ij} \langle f(\mathbf{Y}_i), f(\mathbf{Y}_j) \rangle \quad (3.2.22)$$

and

$$\|\mathbf{y}_{n+1}\|^2 = \sum_{i=1}^s b_i \|\mathbf{Y}_i\|^2 + 2h \sum_{i,j=1}^s b_i e_{ij} \langle \mathbf{Y}_i, f(\mathbf{Y}_j) \rangle + h^2 \sum_{i=1}^s b_i \left\| \sum_{j=1}^s e_{ij} f(\mathbf{Y}_j) \right\|^2 \quad (3.2.23)$$

where $e_{ij} = b_j - a_{ij}$ and m_{ij} is defined by (3.2.19).

Proof. To establish (3.2.22) note that by (3.2.2)

$$\|\mathbf{y}_{n+1}\|^2 = \|\mathbf{y}_n\|^2 + 2h \sum_{i=1}^s b_i \langle \mathbf{y}_n, f(\mathbf{Y}_i) \rangle + h^2 \sum_{i,j=1}^s b_i b_j \langle f(\mathbf{Y}_i), f(\mathbf{Y}_j) \rangle. \quad (3.2.24)$$

Now (3.2.1) implies

$$\mathbf{y}_n = \mathbf{Y}_i - h \sum_{j=1}^s a_{ij} f(\mathbf{Y}_j)$$

Hence

$$\langle \mathbf{y}_n, f(\mathbf{Y}_i) \rangle = \langle \mathbf{Y}_i, f(\mathbf{Y}_i) \rangle - h \sum_{j=1}^s a_{ij} \langle f(\mathbf{Y}_j), f(\mathbf{Y}_i) \rangle,$$

and substituting for $\langle \mathbf{y}_n, f(\mathbf{Y}_i) \rangle$ in (3.2.24) implies (3.2.22). To establish (3.2.23) subtract (3.2.1) from (3.2.2) which yields

$$\mathbf{y}_{n+1} = \mathbf{Y}_i + h \sum_{j=1}^s e_{ij} f(\mathbf{Y}_j),$$

and taking norms of both sides gives

$$\|\mathbf{y}_{n+1}\|^2 = \|\mathbf{Y}_i\|^2 + 2h \sum_{j=1}^s e_{ij} \langle \mathbf{Y}_i, \mathbf{f}(\mathbf{Y}_j) \rangle + h^2 \left\| \sum_{j=1}^s e_{ij} \mathbf{f}(\mathbf{Y}_j) \right\|^2.$$

Recalling that $\sum_{i=1}^s b_i = 1$, multiply both sides by b_i and sum over i to obtain the result. ■

3.3 Linear Multistep and One-Leg Methods

A general k -step linear multistep method for approximating the solution of (2.1.1) may be written as

$$\sum_{j=0}^k \alpha_j \mathbf{y}_{n+j} = h \sum_{j=0}^k \beta_j \mathbf{f}(\mathbf{y}_{n+j}). \quad (3.3.1)$$

Here \mathbf{y}_{n+j} approximates the exact solution $\mathbf{y}(t_{n+j})$ at $t_{n+j} = (n+j)h$, where $h > 0$ is the fixed step-size. The parameters α_i and β_i define the particular method.

Define the polynomials $\rho(z)$, $\sigma(z)$ by

$$\rho(z) = \sum_{j=0}^k \alpha_j z^j, \quad \sigma(z) = \sum_{j=0}^k \beta_j z^j \quad (3.3.2)$$

and letting E be the translation operator $E\mathbf{y}_n = \mathbf{y}_{n+1}$ then (3.3.1) can be rewritten in more compact notation as

$$\rho(E)\mathbf{y}_n = h\sigma(E)\mathbf{f}(\mathbf{y}_n). \quad (3.3.3)$$

We assume throughout that the method (3.3.3) is consistent and zero-stable. This implies that

$$\rho(1) = 0, \quad \text{and} \quad \sigma(1) = \rho'(1) \neq 0,$$

and without loss of generality we can further assume that

$$\sigma(1) = 1.$$

If $\beta_k = 0$ then the method is *explicit*, and it is *implicit* otherwise. If ρ and σ have no common factors then the method is said to be *irreducible*, and otherwise it is said to be *reducible*. Suppose (3.3.3) is reducible and d is a common factor of ρ and σ , then

let

$$\tilde{\rho}(z) = \frac{\rho(z)}{d(z)}, \quad \tilde{\sigma}(z) = \frac{\sigma(z)}{d(z)}. \quad (3.3.4)$$

Now consider the reduced method

$$\tilde{\rho}(E)\mathbf{y}_n = h\tilde{\sigma}(E)\mathbf{f}(\mathbf{y}_n). \quad (3.3.5)$$

Multiplication by $d(E)$ shows that any solution of the simpler method (3.3.5) is also a solution of (3.3.3), therefore it is usual to restrict attention to irreducible methods. (Note however that (3.3.3) and (3.3.5) are not equivalent as (3.3.3) may admit solutions that do not satisfy (3.3.5)).

To every method of the form (3.3.1) there is an associated one-leg method

$$\sum_{j=0}^k \alpha_j \mathbf{x}_{n+j} = h\mathbf{f}\left(\sum_{j=0}^k \beta_j \mathbf{x}_{n+j}\right) \quad (3.3.6)$$

which may be written in compact notation as

$$\rho(E)\mathbf{x}_n = h\mathbf{f}(\sigma(E)\mathbf{x}_n). \quad (3.3.7)$$

Note that, one disadvantage of both linear multistep and one-leg methods is that starting values $\mathbf{y}_0, \dots, \mathbf{y}_{k-1}$ are needed to begin the iteration. We will assume that starting values are given; these could be obtained from \mathbf{y}_0 using a Runge-Kutta method, for example.

Note that the linear multistep method (3.3.3) and the one-leg method (3.3.7) are equivalent for linear problems. The following result of Dahlquist [11], shows that this equivalence runs deeper.

Result 3.3.1 *Given an irreducible linear multistep method (3.3.3) and associated one-leg method (3.3.7) there exist two polynomials P, Q of degree not exceeding $k - 1$ such that*

$$P(\zeta)\sigma(\zeta) - Q(\zeta)\rho(\zeta) = 1. \quad (3.3.8)$$

Now suppose $\{\mathbf{y}_n\}_{n=0}^{\infty}$ satisfies the linear multistep method (3.3.3) and let

$$\mathbf{x}_n = P(E)\mathbf{y}_n - hQ(E)\mathbf{f}(\mathbf{y}_n) \quad (3.3.9)$$

then $\mathbf{y}_n = \sigma(E)\mathbf{x}_n$ and $\{\mathbf{x}_n\}_{n=0}^\infty$ satisfies the one-leg method (3.3.7).

Conversely, suppose $\{\mathbf{x}_n\}_{n=0}^\infty$ satisfies the one-leg method (3.3.7), and let $\mathbf{y}_n = \sigma(E)\mathbf{x}_n$, then $\{\mathbf{y}_n\}_{n=0}^\infty$ satisfies the linear multistep method (3.3.3) and \mathbf{x}_n satisfies (3.3.9). ■

The above results show that linear multistep and one-leg methods are closely related. Although they are rarely used in practice, the behaviour of one-leg methods has been studied in detail. This is because they are often easier to analyse than linear multistep methods and results are easier to formulate in the one-leg setting. Of course, using the equivalence above, it is possible to translate any result for one-leg methods into a related linear multistep method result.

3.4 Numerical Stability Theories

In this section we give a very brief review of numerical stability theories. A complete account can be found in [26] and [27], amongst other places.

Dahlquist [10] introduced the now classic complex linear problem test problem, find $y \in \mathbb{C}$ such that

$$\frac{dy}{dt} = \lambda y \quad \text{for } t \geq 0 \quad \text{and} \quad y(0) = y_0 \quad (3.4.1)$$

where $\lambda \in \mathbb{C}$, and introduced the concept of A-stability.

Definition 3.4.1 A numerical method is *A-stable* if for any $y_0 \in \mathbb{C}$, any fixed step-size h , and any λ such that $\operatorname{Re}(\lambda) < 0$ it follows that the numerical solution of (3.4.1) satisfies $|y_n| \rightarrow 0$ as $n \rightarrow \infty$.

Remark A-stability is more usually defined by requiring that $|y_n|$ remains bounded for any λ such that $\operatorname{Re}(\lambda) \leq 0$, however using the Maximum Modulus principle it can be shown that the definition given above is equivalent to the more usual definition.

For a Runge-Kutta method it can be shown (see [13, 27]) that the numerical solution of (3.4.1) is given by

$$\mathbf{y}_{n+1} = R(h\lambda)\mathbf{y}_n \quad (3.4.2)$$

and hence

$$\mathbf{y}_n = R(h\lambda)^n \mathbf{y}_0 \quad (3.4.3)$$

where the *stability function* $R(z)$ is defined by

$$R(z) = 1 + z\mathbf{b}^T(I - zA)^{-1}\mathbf{1} \quad (3.4.4)$$

and $\mathbf{1} = (1, 1, \dots, 1)^T \in \mathbb{R}^m$. Thus a Runge-Kutta method is A-stable if $|R(z)| \leq 1$ for all z such that $\operatorname{Re}(z) \leq 0$, and to determine whether a Runge-Kutta method is A-stable we need only calculate its stability function.

Dahlquist [11], was also the first to derive a general nonlinear stability theory of ordinary differential equations. He considered the numerical solution of the nonautonomous equivalent to (2.1.1, 2.1.8) using irreducible one-leg methods and introduced the concept of G-stability for these methods. The definition of G-stability is somewhat involved, and since we will not refer to it again we do not reproduce it here. Roughly speaking G-stability implies that a weighted norm over k steps of the difference of two numerical solution sequences is nonincreasing for any system of the form (2.1.1, 2.1.8).

Dahlquist [12] was able to classify the G-stable irreducible one-leg methods and proved the remarkable result that G-stability is equivalent to A-stability.

G-stability theory is specific to irreducible one-leg methods. Butcher [8] considered the solution of (2.1.1, 2.1.8) by Runge-Kutta methods and introduced the concept of B-stability.

Definition 3.4.2 The Runge-Kutta method (3.2.1–2) is said to be *B-stable* if any two solution sequences $\{\mathbf{y}_n\}_{n=0}^\infty$ and $\{\mathbf{y}'_n\}_{n=0}^\infty$ of the method applied to (2.1.1, 2.1.8) satisfy $\|\mathbf{y}_{n+1} - \mathbf{y}'_{n+1}\| \leq \|\mathbf{y}_n - \mathbf{y}'_n\|$ for all $n \geq 1$.

In order to classify the B-stable Runge-Kutta methods Burrage and Butcher [6] later introduced the concept of algebraic stability (recall Definition 3.2.1).

Much effort has been applied to comparing the stability definitions for Runge-Kutta methods, and to classifying the methods which are stable in each sense. This work is well presented in Dekker & Verwer [13] and Hairer & Wanner [27]. Recalling the concept of S-irreducibility from Section 3.2.1 we present without proof:

Result 3.4.3 *For S-irreducible Runge-Kutta methods*

$$\text{algebraic stability} \Leftrightarrow \text{B-stability} \Rightarrow \text{A-stability}.$$

The implications to the right hold for all Runge-Kutta methods. ■

Note that the two-stage theta method (3.5.2), defined below, is A-stable for $\theta \in [1/2, 1]$ but is only algebraically stable for $\theta = 1$. Hence, unlike the situation with one-leg methods, B-stability and A-stability are not equivalent for Runge-Kutta methods, even if we restrict attention to S-irreducible methods.

Although it is just an algebraic condition on the parameters of the method, algebraic stability, and in particular the M matrix, plays an important role in determining the dynamics of the numerical solution. Methods which are algebraically stable or satisfy $M = 0$ have been seen to preserve qualitative features of the underlying flow for a variety of classes of system. In Chapter 6 we will show that if a dissipative system of the form (2.1.1, 2.2.1) solved numerically using an algebraically stable Runge-Kutta method then the numerical approximation preserves the dissipativity of the underlying system. In contrast for Hamiltonian systems, which have no dissipation, Sanz-Serna [47] has proved that Runge-Kutta methods which satisfy $M = 0$ preserve the symplectic structure of the flow.

Although many positive results can be proved for algebraically stable methods, it should be noted that all such methods are implicit, which is a major drawback in their implementation. However we have recently extended the concept of algebraic stability to variable time-stepping methods where explicit methods can be “essentially algebraically stable” and preserve such features as the dissipativity of the underlying flow [55].

3.5 Examples

We now give examples of Runge-Kutta, linear multistep and one-leg methods. The simplest Runge-Kutta methods are the one-stage theta methods, which using Butcher tableau notation may be written as

$$\begin{array}{c|c} \theta & \theta \\ \hline & 1 \end{array} \quad (3.5.1)$$

where $\theta \in [0, 1]$. This method may also be written as

$$\mathbf{y}_{n+1} = \mathbf{y}_n + hf((1 - \theta)\mathbf{y}_n + \theta\mathbf{y}_{n+1}),$$

and in this form we see that the one-stage theta method is also the general consistent one-step one-leg method.

Another simple class of Runge-Kutta methods are the two-stage theta methods

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1-\theta & \theta \\ \hline & 1-\theta & \theta \end{array} \quad (3.5.2)$$

where $\theta \in [0, 1]$. This method may also be written as

$$\mathbf{x}_{n+1} = \mathbf{x}_n + h \left[(1-\theta) \mathbf{f}(\mathbf{x}_n) + \theta \mathbf{f}(\mathbf{x}_{n+1}) \right].$$

and in this form we see that this method is also the general consistent one-step linear multistep method.

Example 3.5.1 Note that by Result 3.3.1 the solution sequences of the one- and two-stage theta methods are related. For both methods

$$\rho(z) = z - 1, \quad \sigma(z) = (1-\theta) + \theta z, \quad (3.5.3)$$

and (3.3.8) is satisfied with $P \equiv 1$ and $Q \equiv \theta$. Thus if we fix $\theta \in [0, 1]$, suppose that $\{\mathbf{y}_n\}_{n=0}^{\infty}$ satisfies the two-stage theta method (3.5.2), and let $\mathbf{x}_n = \mathbf{y}_n - h\theta \mathbf{f}(\mathbf{y}_n)$, then $\mathbf{y}_n = (1-\theta)\mathbf{x}_n + \theta\mathbf{x}_{n+1}$ and $\{\mathbf{x}_n\}_{n=0}^{\infty}$ satisfies the one-stage theta method (3.5.1) for $n \geq 0$.

Conversely suppose that $\{\mathbf{x}_n\}_{n=0}^{\infty}$ satisfies the one-stage theta method (3.5.1), and let $\mathbf{y}_n = (1-\theta)\mathbf{x}_n + \theta\mathbf{x}_{n+1}$, then $\{\mathbf{y}_n\}_{n=0}^{\infty}$ satisfies the two-stage theta method (3.5.2) and $\mathbf{x}_n = \mathbf{y}_n - h\theta \mathbf{f}(\mathbf{y}_n)$. \square

Note that both the one- and two-stage theta methods reduce to the forward Euler method (3.1.1) when $\theta = 0$. The backward Euler method

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}(\mathbf{y}_{n+1}), \quad (3.5.4)$$

corresponds to taking $\theta = 1$ in either (3.5.1) or (3.5.2).

Although the solutions of (3.5.1) and (3.5.2) are closely related it should be noted that whilst both methods are A-stable for $\theta \in [1/2, 1]$ and the one-stage theta method

(3.5.1) is also algebraically stable for $\theta \in [1/2, 1]$, the two-stage theta method (3.5.2) is only algebraically stable if $\theta = 1$.

Setting $\theta = 1/2$ in (3.5.1) gives the implicit midpoint rule. This is the unique second order one-stage method. This method belongs to a class of methods called the Butcher IA methods, which are based on Gauss-Legendre quadrature. They were first tabulated by Butcher [7] in 1964. The two-stage method in this class is

$$\begin{array}{c|cc} \frac{1}{2} - \frac{1}{6}\sqrt{3} & \frac{1}{4} & \frac{1}{4} - \frac{1}{6}\sqrt{3} \\ \frac{1}{2} + \frac{1}{6}\sqrt{3} & \frac{1}{4} + \frac{1}{6}\sqrt{3} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \quad (3.5.5)$$

Methods can be derived in this class with arbitrarily many stages, with an s stage method having order $2s$. The methods in this class all have $M = 0$ (where M is defined by (3.2.12)) and B positive definite, and hence are algebraically stable. Thus there exist algebraically stable Runge-Kutta methods of arbitrarily high order.

Setting $\theta = 1/2$ in (3.5.2) yields the trapezoidal rule. This method can be obtained by Lobatto quadrature, and belongs to the Lobatto IIIA class of methods. The three stage method in this class is

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/2 & 5/24 & 1/3 & -1/24 \\ 1 & 1/6 & 2/3 & 1/6 \\ \hline 1 & 1/6 & 2/3 & 1/6 \end{array} \quad (3.5.6)$$

Lobatto IIIA methods can be derived with arbitrary many stages, with an s -stage method having order $2s - 2$. Unlike the Butcher IA methods, M is not positive definite for the Lobatto IIIA methods and hence they are not algebraically stable, although they are A-stable.

Other classes of Runge-Kutta methods based on quadrature formulae are the Radau IA, Radau IIA, Lobatto IIIB and Lobatto IIIC methods. All of these classes of methods are algebraically stable except for the Lobatto IIIB methods. The low order methods in these classes can be found tabulated in [13] or [27].

Although the one- and two-stage theta methods are very simple numerical methods we will pay considerable attention to these methods throughout the thesis. When (2.1.1) is derived from a spatial discretization of a partial differential equation, m will

in general be very large. In such a case, due to limitations in storage and/or computing power, it is often necessary to solve (2.1.1) using simple methods, and the two-stage theta method in particular is widely used in practice for this reason. Thus since these methods are popular it is important to study their dynamics, and we will do this.

3.6 Runge-Kutta Methods as Dynamical Systems

A Runge-Kutta method, applied to (2.1.1), not only defines an approximation to the solution of (2.1.1), but can also define a discrete dynamical system in a natural way. In the following chapters we will compare the dynamics of the underlying dynamical system with the dynamics of the discrete dynamical system defined by the discretization. This will allow us to compare and contrast the asymptotic behaviour of the underlying dynamical system with the asymptotic behaviour of its numerical discretization.

The numerical approximation to a given dynamical system of the form (2.1.1) generated by the forward Euler method (3.1.1) defines a discrete dynamical system (of the form (2.5.1)) on \mathbb{R}^m with $\Phi_h(\mathbf{y}) = \mathbf{y} + h\mathbf{f}(\mathbf{y})$.

Similarly the numerical solution generated by any explicit Runge-Kutta method defines a discrete dynamical system with

$$\Phi_h(\mathbf{y}_n) = \mathbf{y}_n + h \sum_{i=1}^s b_i \mathbf{f}(\mathbf{Y}_i),$$

on noting that the stage values \mathbf{Y}_i are uniquely determined at each step for an explicit method. However for an implicit method the Runge-Kutta defining equations (3.2.1) need not be uniquely soluble, and hence an implicit Runge-Kutta method need not define a discrete dynamical system. Thus before we can compare the discrete dynamical system defined by the numerical approximation to (2.1.1) with the underlying dynamical system itself we need to determine under what conditions the Runge-Kutta defining equations (2.1.1) are uniquely soluble. In the following subsections we will consider this problem under continuity and structural assumptions on \mathbf{f} .

3.6.1 Solubility of Implicit Runge-Kutta Equations under Continuity Conditions

If we are to implement implicit Runge-Kutta methods then we need to consider whether (3.2.1) is soluble and if so whether the solution is unique, and how we can compute

the solution. In this section we will consider the solution of (3.2.1) under continuity conditions on \mathbf{f} . This problem was first addressed by Butcher who proved the following result for the case where \mathbf{f} is globally Lipschitz. It implies that the numerical approximation to (2.1.1) by an implicit Runge-Kutta method defines a discrete dynamical system when \mathbf{f} is globally Lipschitz if h is sufficiently small.

Result 3.6.1 (Butcher [7]) *If \mathbf{f} is globally Lipschitz with Lipschitz constant L and*

$$h < \frac{1}{La}, \quad (3.6.1)$$

where a is defined by (3.2.6), then the equations (3.2.1–2) are uniquely soluble. Furthermore this solution can be found by iteration; set $\mathbf{Y}_i^0 = \mathbf{y}_n \quad \forall i = 1, \dots, s$ then iterate

$$\mathbf{Y}_i^{N+1} = \mathbf{y}_n + h \sum_{j=1}^{i-1} a_{ij} \mathbf{f}(\mathbf{Y}_j^{N+1}) + h \sum_{j=i}^s a_{ij} \mathbf{f}(\mathbf{Y}_j^N) \quad (3.6.2)$$

and let $\mathbf{Y}_i = \lim_{N \rightarrow \infty} \mathbf{Y}_i^N$. This limit exists and defines the solution of (3.2.1). ■

The existence of a (not necessarily unique) solution of the defining equations can be proved under the much weaker assumption that \mathbf{f} is continuous.

Proposition 3.6.2 *Suppose \mathbf{f} is continuous on $\mathcal{N}(B, \varepsilon)$ and that $M < \infty$ where B is some subset of \mathbb{R}^m , $\varepsilon > 0$ and*

$$M = \sup_{\mathbf{y} \in \mathcal{N}(B, \varepsilon)} \|\mathbf{f}(\mathbf{y})\|. \quad (3.6.3)$$

Then if

$$h < \frac{\varepsilon}{aM} \quad (3.6.4)$$

(where a is defined by (3.2.6)) and $\mathbf{y}_n \in B$ there exists a solution of the equations (3.2.1–2) such that

$$\|\mathbf{Y}_i - \mathbf{y}_n\| < \varepsilon \quad \forall i \quad (3.6.5)$$

and hence $\mathbf{Y}_i \in B(\mathbf{y}_n, \varepsilon) \subseteq \mathcal{N}(B, \varepsilon) \quad \forall i$. Furthermore, if the iteration (3.6.2) converges, then it converges to such a solution.

Proof. By (3.6.4) there exists $\varepsilon^* \in (0, \varepsilon)$ such that

$$h \leq \frac{\varepsilon^*}{aM}. \quad (3.6.6)$$

Now consider the iteration (3.6.2) and denote the Cartesian product of s closed balls

Now consider the iteration (3.6.2) and denote the Cartesian product of s closed balls $\overline{B}(\mathbf{y}_n, \varepsilon^*)$, by $\overline{B}(\mathbf{y}_n, \varepsilon^*)^s$. Suppose

$$(\mathbf{Y}_1^N, \mathbf{Y}_2^N, \dots, \mathbf{Y}_s^N) \in B(\mathbf{y}_n, \varepsilon^*)^s$$

then (3.6.2) and (3.6.6) imply that

$$(\mathbf{Y}_1^{N+1}, \mathbf{Y}_2^{N+1}, \dots, \mathbf{Y}_s^{N+1}) \in B(\mathbf{y}_n, \varepsilon^*)^s.$$

Furthermore, since \mathbf{f} is continuous on $\mathcal{N}(B, \varepsilon)$ the iteration (3.6.2) defines a continuous map from the convex compact set $\overline{B}(\mathbf{y}_n, \varepsilon^*)^s$ into itself. Thus by Result 2.6.1 there exists a fixed point of the iteration (3.6.2) within $\overline{B}(\mathbf{y}_n, \varepsilon^*)^s$. This defines the required solution of (3.2.1–2), and clearly if the iteration (3.6.2) converges then it must converge to such a solution. ■

Remark We have not proved any of the following:

- (i) that there is a unique solution of (3.2.1–2) satisfying the properties given in Proposition 3.6.2,
- (ii) that there is not a solution of (3.2.1–2) such that $\|\mathbf{Y}_i - \mathbf{y}_n\| > \varepsilon$ for some (or all) i ,
- (iii) that the iteration (3.6.2) converges.

If we assume that \mathbf{f} is Lipschitz on $\mathcal{N}(B, \varepsilon)$ then we can prove that (i) and (iii) hold.

Proposition 3.6.3 Suppose \mathbf{f} is Lipschitz on $\mathcal{N}(B, \varepsilon)$ and that $M < \infty$ where B is some subset of \mathbb{R}^m , $\varepsilon > 0$ and M is defined by (3.6.3). If

$$h < \min \left(\frac{\varepsilon}{aM}, \frac{1}{La} \right) \quad (3.6.7)$$

where a is defined by (3.2.6), then for any $\mathbf{y}_n \in B$ there exists a unique solution of the equations (3.2.1–2), such that

$$\|\mathbf{Y}_i - \mathbf{y}_n\| < \varepsilon \quad \forall i$$

and hence $\mathbf{Y}_i \in B(\mathbf{y}_n, \varepsilon) \subseteq \mathcal{N}(B, \varepsilon) \forall i$, and the iteration (3.6.2) converges to this solution.

Proof. By the proof of Proposition 3.6.2

$$\|Y_j^N - y_n\| \leq \varepsilon^* \quad \forall j, N.$$

So $Y_j^N \in U \quad \forall j, N$. Now f is Lipschitz on $\mathcal{N}(B, \varepsilon)$ and although Result 3.6.1 does not apply in this case, Butcher's proof [7] holds, to give the required result. ■

We will often consider the numerical approximation of (2.1.1) where f satisfies some structural assumption. Sometimes the structure imposed on f will imply an *a priori* bound on the numerical solution, and in this case the following theorem shows that any implicit Runge-Kutta method will define a discrete dynamical system if the step-size is sufficiently small.

Theorem 3.6.4 *The numerical solution generated by the Runge-Kutta method (3.2.1–2) defines a continuous discrete dynamical system on a set $B \subseteq \mathbb{R}^m$ if*

- (i) *f is Lipschitz on $\mathcal{N}(B, \varepsilon)$ with Lipschitz constant L for some $\varepsilon > 0$,*
- (ii) *h satisfies (3.6.7),*
- (iii) *for an implicit method the solution of (3.2.1–2) is defined by Proposition 3.6.3, and,*
- (iv) *there is an a priori bound which implies that if $y_n \in B$ then the solution of (3.2.1–2) satisfies $y_{n+1} \in B$.*

Proof. Given $y_n \in B$, Proposition 3.6.3 defines a unique solution of the Runge-Kutta defining equations (3.2.1–2) with $Y_i \in \mathcal{N}(B, \varepsilon)$ for all i , and then the *a priori* bound on the solution implies that $y_{n+1} \in B$, and hence that the numerical solution defines a discrete dynamical system on B . It only remains to show that this system is continuous with respect to initial data. To establish this let

$$Z_i = z_n + h \sum_{j=1}^s a_{ij} f(Z_j), \quad i = 1, \dots, s$$

and

$$z_{n+1} = z_n + h \sum_{i=1}^s b_i f(Z_i).$$

Then

$$Y_i - Z_i = (y_n - z_n) + h \sum_{j=1}^s a_{ij} [f(Y_j) - f(Z_j)]$$

and letting

$$M = \max_{1 \leq j \leq s} \|Y_j - Z_j\|$$

we obtain

$$\begin{aligned} \|Y_i - Z_i\| &\leq \|y_n - z_n\| + hA \max_{1 \leq j \leq s} \|f(Y_j) - f(Z_j)\| \\ &\leq \|y_n - z_n\| + LAhM. \end{aligned} \quad (3.6.8)$$

But (3.6.8) holds for all i and hence noting that (3.6.7) implies that $h < 1/LA$ it follows that

$$\begin{aligned} M &\leq \|y_n - z_n\| + LAhM \\ &\leq \frac{1}{1 - LAh} \|y_n - z_n\| \end{aligned}$$

Furthermore,

$$\begin{aligned} \|y_{n+1} - z_{n+1}\| &\leq \|y_n - z_n\| + h \sum_{j=1}^s |b_j| \|f(Y_j) - f(Z_j)\| \\ &\leq \|y_n - z_n\| + hBLM \\ &\leq \frac{1 + Lh(\mathbf{B} - \mathbf{A})}{1 - LAh} \|y_n - z_n\|, \end{aligned}$$

which proves continuity with respect to initial data. ■

3.6.2 Local and Global Error Bounds

In this section we will derive error bounds for the numerical solution when f is Lipschitz but not necessarily differentiable. We will use these bounds in Chapter 7 to prove that the numerical solution possesses an attractor close to the global attractor of the underlying dissipative system that we are approximating numerically.

Local and global error bounds for Runge-Kutta methods can be found in most good numerical analysis books including Hairer, Nørsett and Wanner [26] and Butcher [9], however the approach usually adopted is to assume that f is p times differentiable and then to show that a method of order p has local truncation error of order $p + 1$ and global truncation error of order p . To make our results applicable to as wide a class of problems as possible we will assume the weakest differentiability conditions that allow us to prove our results; usually this will mean that f is locally Lipschitz, and so we

derive error bounds under this assumption. By treating all Runge-Kutta methods as perturbations of the forward Euler method we show that they all have local truncation error $O(h^2)$ and global truncation error $O(h)$.

Definition 3.6.5 The *local truncation error* $l(h, \mathbf{y})$ of the Runge-Kutta method (3.2.1–2) is defined to be the norm of the error in the numerical solution over one step of size h with initial condition \mathbf{y} , so

$$l(h, \mathbf{y}_n) = \|S(h)\mathbf{y}_n - \mathbf{y}_{n+1}\|. \quad (3.6.9)$$

Note that for an implicit method this will not be well defined, unless the defining equations (3.2.1–2) are uniquely soluble. However, in Proposition 3.6.3 we showed the existence of a locally unique solution of the defining equations (3.2.1–2), and assuming that the numerical solution is defined by this solution of (3.2.1–2) the local truncation error is well defined for implicit methods.

We first derive a bound on the local truncation error of the forward Euler method (3.1.1).

Lemma 3.6.6 Suppose that (2.1.1) defines a dynamical system on \mathbb{R}^m , \mathbf{f} is Lipschitz with Lipschitz constant L on a set $B \subseteq \mathbb{R}^m$, and that $S(t)\mathbf{y}_0 \in B$ for $t \in [0, h]$, then the local truncation error of the forward Euler method satisfies

$$l(h, \mathbf{y}_0) \leq \frac{1}{2}h^2 Le^{Lh} \|\mathbf{f}(\mathbf{y}_0)\|. \quad (3.6.10)$$

Proof. For $t \in [0, h]$

$$\begin{aligned} 2l(t, \mathbf{y}_0) \frac{d}{dt} l(t, \mathbf{y}_0) &= \frac{d}{dt} l(t, \mathbf{y}_0)^2 = 2\langle S(t)\mathbf{y}_0 - \mathbf{y}_0 - t\mathbf{f}(\mathbf{y}_0), \mathbf{f}(\mathbf{y}(t)) - \mathbf{f}(\mathbf{y}_0) \rangle \\ &\leq 2l(t, \mathbf{y}_0) \|\mathbf{f}(\mathbf{y}(t)) - \mathbf{f}(\mathbf{y}_0)\| \end{aligned}$$

and hence

$$\frac{d}{dt} l(t, \mathbf{y}_0) \leq \|\mathbf{f}(\mathbf{y}(t)) - \mathbf{f}(\mathbf{y}_0)\|$$

if $l(t, \mathbf{y}_0) \neq 0$. Thus by Lipschitz continuity and Lemma 2.1.5

$$\frac{d}{dt} l(t, \mathbf{y}_0) \leq L\|\mathbf{y}(t) - \mathbf{y}_0\|$$

$$\leq \|f(y_0)\| [e^{Lt} - 1]$$

when $l(t, y_0) \neq 0$. Hence

$$\begin{aligned} l(h, y_0) &\leq \|f(y_0)\| \int_0^h [e^{Lt} - 1] dt \\ &= \frac{1}{L} \|f(y_0)\| [e^{Lt} - Lt]_{t=0}^h \\ &= \frac{1}{L} \|f(y_0)\| [e^{Lh} - 1 - Lh]. \end{aligned}$$

Now the result follows on noting that

$$\begin{aligned} e^{Lh} - 1 - Lh &= h^2 L^2 \sum_{k=2}^{\infty} \frac{(hL)^{k-2}}{k!} \\ &\leq \frac{h^2 L^2}{2} e^{hL}. \blacksquare \end{aligned}$$

So we have shown that the local truncation error (3.6.10) of the forward Euler method is still $O(h^2)$ when f is Lipschitz continuous but not necessarily differentiable. By treating all other Runge-Kutta methods as perturbations of the forward Euler method we now use Lemma 3.6.6 to show that the local truncation error of any Runge-Kutta method is at worst $O(h^2)$ when f is Lipschitz continuous.

Proposition 3.6.7 *Suppose (2.1.1) defines a dynamical system on \mathbb{R}^m , and that $S(t)y_0 \in B \forall t \in [0, h]$ for some subset $B \subseteq \mathbb{R}^m$. If for some $\varepsilon > 0$, f is Lipschitz on $\mathcal{N}(B, \varepsilon)$ with Lipschitz constant L , M defined by*

$$M = \sup_{y \in \mathcal{N}(B, \varepsilon)} \|f(y)\| \quad (3.6.11)$$

is finite and (3.6.7) is satisfied then the local truncation error of the Runge-Kutta method (3.2.1-2) (where for an implicit method the solution is defined by Proposition 3.6.3) satisfies

$$l(h, y_0) \leq h^2 L \|f(y_0)\| \left[\frac{1}{2} e^{Lh} + \frac{AB}{1 - LAh} \right] \quad (3.6.12)$$

and hence given any $C > \frac{1}{2} + \mathbf{AB}$ there exists $h(C) > 0$ such that for $h \in (0, h(C))$

$$l(h, \mathbf{y}_0) \leq Ch^2 LM. \quad (3.6.13)$$

Proof. By Proposition 3.6.3 the Runge-Kutta defining equations (3.2.1) are soluble. Now consider

$$\begin{aligned} l(h, \mathbf{y}_0) &= \|S(h)\mathbf{y}_0 - \mathbf{y}_1\| \\ &= \|S(h)\mathbf{y}_0 - (\mathbf{y}_0 + h \sum_{i=1}^s b_i \mathbf{f}(\mathbf{Y}_i))\| \\ &= \|S(h)\mathbf{y}_0 - (\mathbf{y}_0 + h\mathbf{f}(\mathbf{y}_0)) - h \sum_{i=1}^s b_i (\mathbf{f}(\mathbf{Y}_i) - \mathbf{f}(\mathbf{y}_0))\| \\ &\leq \|S(h)\mathbf{y}_0 - (\mathbf{y}_0 + h\mathbf{f}(\mathbf{y}_0))\| + h \sum_{i=1}^s |b_i| \|\mathbf{f}(\mathbf{Y}_i) - \mathbf{f}(\mathbf{y}_0)\| \end{aligned}$$

and hence, denoting the local truncation error of the forward Euler method by $L(h, \mathbf{y}_0)$ we have that

$$l(h, \mathbf{y}_0) = L(h, \mathbf{y}_0) + h \sum_{i=1}^s |b_i| \|\mathbf{f}(\mathbf{Y}_i) - \mathbf{f}(\mathbf{y}_0)\|.$$

Note that $a \geq \mathbf{A}$ and hence (3.6.7) implies that $h < 1/L\mathbf{A}$. Now applying Lemmas 3.6.6 and 3.2.3 gives (3.6.12).

Finally, to derive (3.6.13), let

$$g(h) = \frac{1}{2}e^{Lh} + \frac{\mathbf{AB}}{1 - L\mathbf{A}h}.$$

Note that $g(0) = \frac{1}{2} + \mathbf{AB}$, $g(h)$ is a strictly increasing function of h for $h < 1/L\mathbf{A}$ and that $g(h) \rightarrow \infty$ as $h \rightarrow 1/L\mathbf{A}$. Thus there exists a unique $h(C) > 0$ such that $g(h(C)) = C$, and then $g(h) \leq C$ for $h \in (0, h(C))$. Now reducing $h(C)$ if necessary so that (3.6.7) holds for $h \in (0, h(C))$ the result follows from (3.6.12). ■

Remark We have shown that the local truncation error of all Runge-Kutta methods is at worst $O(h^2)$ when \mathbf{f} is Lipschitz continuous. For methods which are higher than first order we would usually expect the local truncation error to be $O(h^3)$ or higher, but to derive such bounds we must impose differentiability conditions on \mathbf{f} . However we do not assume that \mathbf{f} is differentiable, since we want our results to apply to as wide a class of dynamical systems as possible, and thus the error bound given in Proposition 3.6.7 will not be optimal if \mathbf{f} is differentiable and a second or higher order method is used.

Moreover our proof of the convergence of the numerical attractor to the attractor of the underlying system in Chapter 7, for which the results of this section are needed, does not give a rate of convergence and so the nonoptimality of the bounds in this section when \mathbf{f} is differentiable does not degrade our later results.

Proposition 3.6.7 allows us to derive a bound on the global error of the numerical solution.

Proposition 3.6.8 *Suppose (2.1.1) defines a dynamical system on \mathbb{R}^m , and that $S(t)\mathbf{y}_0 \in N \ \forall t \in [0, t_0]$ for some subset $N \subseteq \mathbb{R}^m$. If for some $\varepsilon > 0$, \mathbf{f} is Lipschitz on $\mathcal{N}(N, \varepsilon)$ with Lipschitz constant L , and M defined by (3.6.11) is finite then given any $C > \frac{1}{2} + \mathbb{A}\mathbb{B}$ there exists $h(C) > 0$ such that for $h \in (0, h(C))$*

(i) *Proposition 3.6.3 implies the existence of a unique sequence $\{\mathbf{y}_n\}_{n=0}^{n^*}$ and associated stage values which satisfy the Runge-Kutta defining equations (3.2.1–2), where n^* is the largest integer such that $n^*h \leq t_0$,*

(ii) $\mathbf{y}_n \in \mathcal{N}(N, \varepsilon/2) \ \forall n \in \{0, 1, \dots, n^*\}$,

(iii) *the global error*

$$e_n(h) := \|S(nh)\mathbf{y}_0 - S_h^n \mathbf{y}_0\| \quad (3.6.14)$$

satisfies

$$e_{n+1}(h) \leq e^{Lh} e_n(h) + l(h, \mathbf{y}_n) \quad (3.6.15)$$

for $n \leq n^* - 1$, and hence,

(iv) for $nh \in [0, t_0]$

$$e_n(h) \leq CMh(e^{Lt_0} - 1). \quad (3.6.16)$$

Proof. Notice that if

$$e_n(h) \leq Ch^2 LM \left[\frac{e^{Lnh} - 1}{e^{Lh} - 1} \right] \quad (3.6.17)$$

then

$$\begin{aligned} e_n(h) &\leq Ch^2 LM \left[\frac{e^{Lnh} - 1}{Lh} \right] \\ &\leq ChM(e^{Lnh} - 1) \end{aligned}$$

and (3.6.16) follows for $nh \in [0, t_0]$.

We will prove the result by induction. Suppose that $\mathbf{y}_m \in \mathcal{N}(N, \varepsilon/2)$, (3.6.15) holds

for $n \leq m - 1$ and (3.6.17) holds for $n \leq m$. Let

$$h_0 = \min \left(\frac{\varepsilon}{2aM}, \frac{1}{La} \right).$$

If $h < h_0$ then Proposition 3.6.3 defines a unique solution of (3.2.1), which satisfies $\mathbf{Y}_i \in \mathcal{N}(N, \varepsilon) \forall i$. Thus the global error at time $t = (m + 1)h$ satisfies

$$\begin{aligned} e_{m+1}(h) &= \|S((m+1)h)\mathbf{y}_0 - S_h^{m+1}\mathbf{y}_0\| \\ &\leq \|S((m+1)h)\mathbf{y}_0 - S(h)S_h^m\mathbf{y}_0\| + \|S(h)S_h^m\mathbf{y}_0 - S_h^{m+1}\mathbf{y}_0\| \\ &= \|S(h)[S(mh)\mathbf{y}_0 - S_h^m\mathbf{y}_0]\| + \|(S(h) - S_h)\mathbf{y}_m\| \\ &= \|S(h)[S(mh)\mathbf{y}_0 - S_h^m\mathbf{y}_0]\| + l(h, \mathbf{y}_m) \end{aligned}$$

Let

$$h_1 = \frac{1}{L} \ln \left(1 + \frac{\varepsilon L}{2M} \right).$$

Since $\mathbf{y}_m \in \mathcal{N}(N, \varepsilon/2)$ if $h < h_1$ then by Lemma 2.1.5 $S(t)\mathbf{y}_m \in \mathcal{N}(N, \varepsilon) \forall t \in [0, h]$, hence we can apply Result 2.1.4 to derive that

$$\begin{aligned} e_{m+1}(h) &\leq e^{Lh} \|S(mh)\mathbf{y}_0 - S_h^m\mathbf{y}_0\| + l(h, \mathbf{y}_m) \\ &= e^{Lh} e_m(h) + l(h, \mathbf{y}_m) \end{aligned}$$

and hence (3.6.15) holds for $n \leq m$. Now by Proposition 3.6.7 there exists $h_2(C) > 0$ such that for $h < h_2(C)$

$$l(h, \mathbf{y}_m) \leq Ch^2 LM$$

and hence using (3.6.17) with $n = m$

$$\begin{aligned} e_{m+1}(h) &\leq e^{Lh} Ch^2 LM \left[\frac{e^{Lmh} - 1}{e^{Lh} - 1} \right] + Ch^2 LM \\ &= Ch^2 LM \left[\frac{e^{L(m+1)h} - 1}{e^{Lh} - 1} \right] \end{aligned}$$

which completes the inductive step for (3.6.17). Finally if $h < h_3$ where

$$h_3 = \frac{\varepsilon}{2CM(e^{L\tau_0} - 1)}$$

then $e_{m+1}(h) \leq \varepsilon/2$ and hence since $S((m+1)h)\mathbf{y}_0 \in N$ it follows that $\mathbf{y}_{m+1} \in$

$\mathcal{N}(N, \varepsilon/2)$. This completes the inductive step, and the result follows on setting $h(C) = \min(h_0, h_1, h_2(C), h_3)$ and noting that (3.6.15) holds for $n = 0$ and (3.6.17) holds for $n = 1$. ■

3.6.3 Solubility of Implicit Runge-Kutta Equations under Structural Assumptions

The structural assumptions which we impose on f in later chapters will not only have implications for the dynamics of the system, but also effect the solubility of the Runge-Kutta defining equations (3.2.1–2). In this section we consider the solubility of the defining equations under the various structural conditions that we will assume.

One-sided Lipschitz Condition

We begin by considering the solution of (3.2.1–2) when f satisfies a one-sided Lipschitz condition (2.1.6). This problem has already been studied extensively in the literature for the numerical solution of stiff systems. We will summarise the existing theory here, see [13] or [27] for a more complete account. We will require some preliminary results before we can establish the existence of solutions to (3.2.1–2) under a one-sided Lipschitz condition. The following definition is reproduced from [13].

Definition 3.6.9 Let D be a positive diagonal $s \times s$ matrix, so $d_{ii} > 0$ for all i , and A an arbitrary $s \times s$ matrix. Then the function $\Psi_D(A)$ is defined by

$$\Psi_D(A) = \inf_{\xi \neq 0} \frac{\langle DA\xi, \xi \rangle}{\langle D\xi, \xi \rangle}. \quad (3.6.1)$$

Now let \mathcal{D} be the set of positive diagonal $s \times s$ matrices and define $\Psi_0(A)$ by

$$\Psi_0(A) = \sup_{D \in \mathcal{D}} \Psi_D(A). \quad (3.6.2)$$

We will often be interested in the $\Psi_D(A^{-1})$ when A is invertible. The following result, which appears as Corollary 5.1.4 in [13], relates $\Psi_D(A)$ to $\Psi_D(A^{-1})$.

Result 3.6.10 (Dekker and Verwer [13]) *The following equivalences hold.*

$$(i) \Psi_D(A) > 0 \iff \{ A \text{ invertible and } \Psi_D(A^{-1}) > 0 \}$$

- (ii) $\Psi_0(A) > 0 \iff \{ A \text{ invertible and } \Psi_0(A^{-1}) > 0 \}$
 (iii) If A is invertible then $\Psi_D(A) = 0 \iff \Psi_D(A^{-1}) = 0$. ■

Recall from Section 3.2.1 that for a DJ-irreducible algebraically stable Runge-Kutta method $b_i > 0$ for all i , so that B is positive definite and $\Psi_B(A)$ is well defined. We will require the following lemma.

Result 3.6.11 *For a DJ-irreducible algebraically stable Runge-Kutta method*

- (i) if A is singular then $\Psi_B(A) = \Psi_0(A) = 0$, whilst
 (ii) if A is invertible then $\Psi_B(A) \geq 0$, $\Psi_B(A^{-1}) \geq 0$ and $\Psi_0(A^{-1}) \geq 0$.

Proof. Observe that

$$\begin{aligned}
 \langle BA\xi, \xi \rangle &= \xi^T BA\xi \\
 &= \frac{1}{2}\xi^T (BA + A^T B - bb^T)\xi + \frac{1}{2}\xi^T bb^T \xi \\
 &= \frac{1}{2}\xi^T M\xi + \frac{1}{2}(b^T \xi)^2 \\
 &\geq 0,
 \end{aligned}$$

since M is positive semi-definite. Now since $\langle B\xi, \xi \rangle = \sum_{i=1}^s b_i \xi_i^2 > 0$ for $\xi \neq 0$ it follows that $\Psi_B(A) \geq 0$.

Now if A singular on choosing ξ in the null space of A it follows from (3.6.1) that $\Psi_D(A) \leq 0$ for any positive diagonal matrix D . Hence $\Psi_B(A) = \Psi_0(A) = 0$ as required.

If A is invertible then the result follows from Result 3.6.10. ■

The following theorem is the main result on the existence and uniqueness of the solutions to the Runge-Kutta defining equations (3.2.1–2) when f satisfies a one-sided Lipschitz condition (2.1.6), and is obtained by combining Theorems 5.3.9 and 5.3.12 of Dekker and Verwer [13].

Result 3.6.12 (Dekker and Verwer [13]) *If f satisfies a one-sided Lipschitz condition (2.1.6) and the Runge-Kutta matrix A is invertible with*

$$hc < \Psi_0(A^{-1}) \tag{3.6.3}$$

then the Runge-Kutta defining equations (3.2.1–2) have a unique solution. Moreover if A is singular and there exists a positive diagonal matrix D such that $\Psi_D(A) = 0$ then the system has exactly one solution when $c < 0$ for any step-size $h > 0$. ■

In the case where $c < 0$ we have an immediate corollary for algebraically stable methods.

Result 3.6.13 *If f satisfies a one-sided Lipschitz condition (2.1.6) with $c < 0$, and the Runge-Kutta method (3.2.1–2) is algebraically stable and DJ-irreducible then there exists a unique solution of the Runge-Kutta defining equations (3.2.1–2) for any step-size $h > 0$.*

Proof. If A is invertible then by Result 3.6.11 (ii) $\Psi_0(A^{-1}) \geq 0$, and since $hc < 0$, (3.6.3) holds and the result follows from Result 3.6.12. If A is singular then Result 3.6.11 (i) implies that $\Psi_B(A) = 0$ and once again we can apply Result 3.6.12. ■

The case where f satisfies a one-sided Lipschitz condition with $c < 0$ is not really very interesting because, as was noted in Chapter 2, this implies that all the trajectories of (2.1.1) are asymptotic to a unique fixed point.

The more interesting case is where $c > 0$ and some expansion of the trajectories is allowed. Result 3.6.12 then tells us that there is a unique solution to the Runge-Kutta defining equations (3.2.1–2) provided that A is invertible, $\Psi_0(A^{-1}) > 0$ and $h < \frac{1}{c} \Psi_0(A^{-1})$. The following values for $\Psi_0(A^{-1})$ for well known algebraically stable Runge-Kutta methods are reproduced from [27], and give an indication of when Result 3.6.12 can be applied.

Result 3.6.14 *For the Butcher IA methods*

$$\Psi_0(A^{-1}) = \min_{i=1,\dots,s} \frac{1}{2c_i(1-c_i)}. \quad (3.6.4)$$

For the Radau IA methods

$$\Psi_0(A^{-1}) = \begin{cases} 1 & \text{if } s = 1 \\ \frac{1}{2(1-c_2)} & \text{if } s > 1 \end{cases} \quad (3.6.5)$$

For the Radau IIA methods

$$\Psi_0(A^{-1}) = \begin{cases} 1 & \text{if } s = 1 \\ \frac{1}{2c_{s-1}} & \text{if } s > 1 \end{cases} \quad (3.6.6)$$

For the Lobatto IIIC methods

$$\Psi_0(A^{-1}) = \begin{cases} 1 & \text{if } s = 2 \\ 0 & \text{if } s > 2 \end{cases} \quad (3.6.7) \quad \blacksquare$$

Note from (3.6.7) that algebraic stability is not sufficient to ensure that $\Psi_0(A^{-1}) > 0$. So far for the Lobatto IIIC method we only have existence and uniqueness, from Result 3.6.13, for $c < 0$. This has been extended to the case $c = 0$ by Hundsdorfer and Spijker [37] for the three stage method, and by Lui and Kraaijevanger [45] for the general method. However existence and uniqueness has not been shown for $c > 0$.

Finally in this section we consider the solubility of the one and two stage theta methods.

Result 3.6.15 *If f satisfies a one-sided Lipschitz condition (2.1.6) then the equations defining the one-stage theta method (3.5.1) are uniquely soluble*

- (i) *for any step-size $h > 0$ if $c \leq 0$ or $\theta = 0$, and,*
- (ii) *for $h < 1/c\theta$ if $c > 0$ and $\theta \in (0, 1]$.*

Proof. If $\theta = 0$ the method is explicit and the result is trivial. If $\theta > 0$ then $A^{-1} = 1/\theta$, and it follows easily that $\Psi_0(A^{-1}) = 1/\theta$ and hence the result follows from Result 3.6.12.

■

Result 3.6.16 *If f satisfies a one-sided Lipschitz condition (2.1.6) then the equations defining the two-stage theta method (3.5.2) are uniquely soluble*

- (i) *for any step-size $h > 0$ if $c \leq 0$ or $\theta = 0$, and,*
- (ii) *for $h < 1/c\theta$ if $c > 0$ and $\theta \in (0, 1]$.*

Proof. Note from (3.5.2) that Y_1 is determined explicitly and indeed $Y_1 = y_n$; hence we need only consider the solubility of the equation defining Y_2 . But on writing $\tilde{y}_n = y_n + h(1 - \theta)f(y_n)$ this equation becomes

$$Y_2 = \tilde{y}_n + h\theta f(Y_2)$$

which is the same as the equation defining Y_1 for the one-stage theta method and so the result follows from Result 3.6.15. ■

Dissipative Structure

Now we consider solving the Runge-Kutta defining equations (3.2.1–2) when the underlying system is dissipative with f satisfying (2.2.1). The following proposition will

allow us to establish the existence of solutions to (3.2.1–2) for algebraically stable DJ-irreducible methods. The proof uses ideas from the existence and uniqueness theory when f satisfies a one-sided Lipschitz condition, and from Foias *et al* [18] who use a result similar to Result 2.6.2, in a similar way to which we do below, whilst proving the existence of solutions to a discretization of the Kuramoto-Sivashinsky equations.

We require some notation in the proof of Proposition 3.6.17 which will be used in the rest of this section. Let $\Psi_D(\bullet)$ be defined by Definition 3.6.9 as in the previous section. Define $Y \in \mathbb{R}^{ms}$ by

$$Y = [Y_1^T, Y_2^T, \dots, Y_s^T]^T \quad (3.6.8)$$

and $F: \mathbb{R}^{ms} \rightarrow \mathbb{R}^{ms}$ by

$$F(Y) = [f(Y_1)^T, f(Y_2)^T, \dots, f(Y_s)^T]^T. \quad (3.6.9)$$

If D is an $s \times s$ positive semi-definite diagonal matrix define a semi-inner product on \mathbb{R}^{ms} by

$$\langle X, Y \rangle_D = X^T (D \otimes I_m) Y \quad (3.6.10)$$

and the corresponding semi-norm on \mathbb{R}^{ms} by

$$\|Y\|_D^2 = \langle Y, Y \rangle_D = Y^T (D \otimes I_m) Y = \sum_{i=1}^s d_i \|Y_i\|^2 \quad (3.6.11)$$

where $(D \otimes I_m)$ denotes the tensor product of D and I_m , the m -dimensional vector of 1's, and $\|\bullet\|$ denotes a norm on \mathbb{R}^m . Note that if D is positive definite then $\langle \bullet, \bullet \rangle_D$ defines an inner product on \mathbb{R}^{ms} and $\|\bullet\|_D$ defines a norm on \mathbb{R}^{ms} .

We will be particularly interested in the semi-inner product and semi-norm on \mathbb{R}^{ms} induced by the matrix B , defined in (3.2.11), associated with the Runge-Kutta method (3.2.1–2). Note that B is positive definite for a DJ-irreducible algebraically stable Runge-Kutta method, and so in this case $\|\bullet\|_B$ defines a norm on \mathbb{R}^{ms} .

Proposition 3.6.17 *If the Runge-Kutta method (3.2.1–2) is DJ-irreducible, A is invertible, f satisfies (2.2.1) and*

$$\Psi_0(A^{-1}) + h\beta > 0 \quad (3.6.12)$$

where $\Psi_0(A)$ is defined by (3.6.2) then the Runge-Kutta defining equations (3.2.1–2) are soluble.

Proof. Define $\mathbf{y} \in \mathbb{R}^{ms}$ by

$$\mathbf{y} = [\mathbf{y}_n^T, \mathbf{y}_n^T, \dots, \mathbf{y}_n^T]^T$$

and as in [13] let

$$\Phi(\mathbf{Y}) = (A^{-1} \otimes I_m)(\mathbf{Y} - \mathbf{y} - h(A \otimes I_m)\mathbf{F}(\mathbf{Y})) \quad (3.6.13)$$

Equation (3.6.12) implies that there exists $\varepsilon > 0$ such that

$$\Psi_0(A^{-1}) + h\beta \geq \varepsilon$$

and the definition of $\Psi_0(A^{-1})$ then implies the existence of a positive definite diagonal matrix D such that

$$\Psi_D(A^{-1}) + h\beta > 0,$$

and by scaling we can choose D such that $\sum_{i=1}^s d_i = 1$. Using this D we have that

$$\langle \mathbf{Y}, \Phi(\mathbf{Y}) \rangle_D = \mathbf{Y}^T(DA^{-1} \otimes I_m)\mathbf{Y} - \mathbf{Y}^T(DA^{-1} \otimes I_m)\mathbf{y} - h\mathbf{Y}^T(D \otimes I_m)\mathbf{F}(\mathbf{Y}) \quad (3.6.14)$$

Consider the terms on the right-hand side of (3.6.14) individually. For the first term it is known that

$$\mathbf{Y}^T(DA^{-1} \otimes I_m)\mathbf{Y} \geq \Psi_D(A^{-1})\|\mathbf{Y}\|_D^2; \quad (3.6.15)$$

see for example Dekker and Verwer [13] or Hairer and Wanner [27]. To bound the second term consider

$$\begin{aligned} \mathbf{Y}^T(DA^{-1} \otimes I_m)\mathbf{y} &= \langle \mathbf{Y}, (A^{-1} \otimes I_m)\mathbf{y} \rangle_D \\ &\leq \|\mathbf{Y}\|_D \|(A^{-1} \otimes I_m)\mathbf{y}\|_D. \end{aligned} \quad (3.6.16)$$

Finally we bound the last term by using the dissipativity of the system. Recall that $\sum_{i=1}^s d_i = 1$. Then using (2.2.1) we have

$$\begin{aligned} \mathbf{Y}^T(D \otimes I_m)\mathbf{F}(\mathbf{Y}) &= \langle \mathbf{Y}, \mathbf{F}(\mathbf{Y}) \rangle_D \\ &= \sum_{i=1}^s d_i \langle \mathbf{Y}_i, \mathbf{f}(\mathbf{Y}_i) \rangle \end{aligned}$$

$$\begin{aligned}
&\leq \alpha \sum_{i=1}^s d_i - \beta \sum_{i=1}^s d_i \|Y_i\|^2 \\
&= \alpha - \beta \|Y\|_D^2
\end{aligned}$$

Substituting all these inequalities into (3.6.14) implies that

$$\langle Y, \Phi(Y) \rangle_D \geq (\Psi_D(A^{-1}) + \beta h) \|Y\|_D^2 - \|Y\|_D \|(A^{-1} \otimes I_m) y\|_D - \alpha h. \quad (3.6.17)$$

Now note that by assumption the coefficient of $\|Y\|_D^2$ is positive so that for $R > x^*$ where x^* is the unique positive zero of the quadratic

$$q(x) = (\Psi_D(A^{-1}) + \beta h)x^2 - (\|(A^{-1} \otimes I_m) y\|_D)x - \alpha h$$

it follows that

$$\langle Y, \Phi(Y) \rangle_D > 0 \quad (3.6.18)$$

for all $Y \in \partial B$ where B is the ball of radius R in $(\mathbb{R}^{ms}, \|\cdot\|_D)$, centred at the origin.

Thus this ball is forward invariant for the dynamical system defined by

$$\frac{dY}{dt} = -\Phi(Y)$$

and it follows from Result 2.6.2 that there exists $Y \in B$ such that $\Phi(Y) = 0$. Thus for this value of Y

$$Y - y - h(A \otimes I_m)F(Y) = 0$$

which is equivalent to

$$Y_i - y_n - h \sum_{j=1}^s a_{ij} f(Y_j) = 0$$

for all $i = 1, \dots, s$ and hence a solution of (3.2.1). ■

Remark Notice that (3.6.18) holds for $\|Y\|_D \geq R$. Hence $\Phi(Y) \neq 0$ for $\|Y\|_D \geq R$ and any solution of (3.2.1) must satisfy $\|Y\|_D < R$. But this is true for any $R > x^*$ and hence $\|Y\|_D \leq x^*$.

Recall that by Result 3.6.11 a DJ-irreducible algebraically stable Runge-Kutta method with invertible A satisfies $\Psi_B(A^{-1}) \geq 0$. Hence the following theorem follows trivially from Proposition 3.6.17.

Theorem 3.6.18 *If the Runge-Kutta method (3.2.1–2) is DJ-irreducible and algebraically stable with A invertible and f satisfies (2.2.1) with $\beta > 0$ then the defining equations (3.2.1) are soluble for any step-size $h > 0$ and any $y_n \in \mathbb{R}^m$. ■*

Remark For a general Runge-Kutta method suppose that A is invertible but that $\Psi_0(A^{-1}) < 0$. In this case Proposition 3.6.17 implies that if

$$h > -\Psi_0(A^{-1})/\beta$$

then there exists a solution of the Runge-Kutta equations (3.2.1). The existence of solutions for h sufficiently large is a rather curious result, contrary to intuition. It may be true that under the assumption (2.2.1) the Runge-Kutta equations (3.2.1) are soluble for any method with A invertible and any step-size $h > 0$, but our theory is not sufficient to show this.

It should be noted that there exist Runge-Kutta methods of arbitrary high order that satisfy the conditions of Theorem 3.6.18. In particular the Butcher IA, Radau IA, Radau IIA and Lobatto IIIC classes of quadrature methods are all DJ-irreducible and algebraically stable with A invertible, as is the backward Euler method.

Having shown the existence of solutions to the Runge-Kutta defining equations (3.2.1) for dissipative systems defined by (2.1.1, 2.2.1) we would also like to derive a global uniqueness result. This however is not possible; in general the solution of (3.2.1) when f satisfies (2.2.1) need not be unique. To show this, we will consider the backward Euler method in one-dimension and will exhibit an f which satisfies (2.2.1) but for which the backward Euler method can have multiple solutions for h arbitrarily small.

In one dimension the backward Euler method is defined by

$$y_{n+1} = y_n + hf(y_{n+1}). \quad (3.6.19)$$

For a given y_n if $h = 0$ then it is trivial that (3.6.19) is uniquely soluble with $y_{n+1} = y_n$. We can use the implicit function theorem to continue this solution for $h > 0$. Define

$$G(y, h) = y - hf(y) - y_n \quad (3.6.20)$$

then $y_{n+1} = y$ is a solution of (3.6.19) if and only if $G(y, h) = 0$. We know $G(y_n, 0) = 0$ and by the implicit function theorem we can continue this solution in h provided $\frac{\partial G}{\partial y} \neq 0$. Now since

$$\frac{\partial G}{\partial y} = 1 - h \frac{df}{dy}(y)$$

we can extend the solution branch provided $h \frac{df}{dy} \neq 1$. If we were to suppose a global bound on $\frac{df}{dy}$, say

$$\frac{df}{dy}(y) \leq c \quad (3.6.21)$$

for all $y \in \mathbb{R}$ then the implicit function theorem gives the existence of a locally unique solution for $h < 1/c$. In fact the solution branch thus defined must be globally unique since two such branches would have to coincide at $h = 0$ contradicting the local uniqueness.

For differentiable functions on \mathbb{R} the one-sided Lipschitz condition (2.1.6) is equivalent to (3.6.21) and so existence and uniqueness of solutions for the backward Euler method follows in this case.

Equation (2.2.1) however, does not imply an upper bound on $\frac{df}{dy}$ and this allows us to construct an example of a system of the form (2.1.1, 2.2.1) for which the backward Euler method admits multiple solutions for any $h > 0$.

Example 3.6.19 Consider the system

$$\frac{dy}{dt} = f(y)$$

on \mathbb{R} where

$$f(y) = -2y + 2 \sin(y^2) \quad (3.6.22)$$

then

$$\begin{aligned} \langle y, f(y) \rangle &= -2y^2 + 2y \sin(y^2) \\ &\leq -2y^2 + y^2 + [\sin(y^2)]^2 \\ &\leq 1 - y^2 \end{aligned}$$

so f defined by (3.6.22) satisfies (2.2.1). We will show the existence of multiple solutions

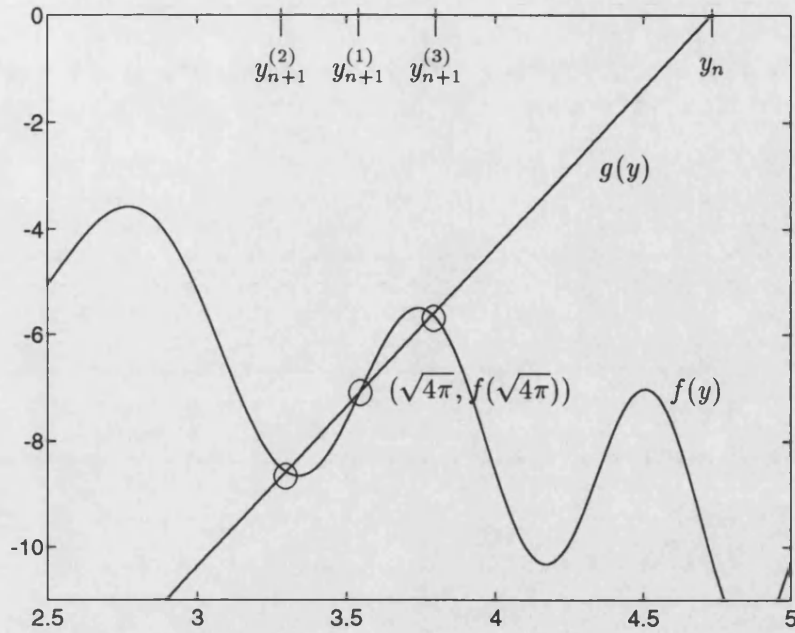


Figure 3.1: Graph of $f(y)$ and $g(y)$ against y with multiple solutions indicated.

of the backward Euler method for arbitrarily small h for this f . Note that

$$\frac{df}{dy}(y) = 4y \cos(y^2) - 2. \quad (3.6.23)$$

Let $y = \sqrt{2k\pi}$ for $k = 1, 2, 3, \dots$ then

$$\frac{df}{dy}(\sqrt{2k\pi}) = 4\sqrt{2k\pi} - 2$$

and since $4\sqrt{2k\pi} - 2 \rightarrow \infty$ as $k \rightarrow \infty$ there is no upper bound on $\frac{df}{dy}$. We construct multiple solutions graphically. Define $g(y)$ by

$$g(y) = f(\sqrt{2k\pi}) + m(y - \sqrt{2k\pi}) \quad (3.6.24)$$

for $m > 0$ and some positive integer k . Plot $f(y)$ and $g(y)$ against y . In the Figure 3.6.19 this is done for $k = 2$. By construction the two lines intersect at $(\sqrt{2k\pi}, f(\sqrt{2k\pi}))$. If we also assume that

$$m < \frac{df}{dy}(\sqrt{2k\pi})$$

then the two lines must also intersect at two other points. Now define y_n to be the unique zero of g and let $h = 1/m$ then (3.6.24) can be rewritten as

$$g(y) = \frac{1}{h}(y - y_n)$$

or

$$y = y_n + hg(y)$$

hence the three intersections of this line with the graph of $f(y)$ define three solutions of the backward Euler method for this h and this y_n . Since $h = 1/m$ we can do this for

$$h \frac{df}{dy}(\sqrt{2k\pi}) > 1$$

which implies

$$h > \frac{1}{4\sqrt{2k\pi} - 2}.$$

Since k is an arbitrary positive integer, given any $h > 0$ we can construct multiple solutions for this step-size by choosing k sufficiently large; however, note that the y_n resulting in multiple solutions satisfy $|y_n| \rightarrow \infty$ as $h \rightarrow 0$. \square

Generalized Dissipative Structure

We now consider the solution of the Runge-Kutta defining equations (3.2.1-2) when f satisfies the generalized dissipativity condition (2.2.9) and prove that if $\Psi_0(A) > 0$ then there exists a solution sequence $\{y_n\}_{n=0}^{\infty}$ for any initial condition y_0 and any step-size $h > 0$.

Theorem 3.6.20 *If the Runge-Kutta method (3.2.1-2) is DJ-irreducible and satisfies*

$$\Psi_0(A) > 0, \tag{3.6.25}$$

where $\Psi_0(A)$ is defined by (3.6.2), and f satisfies (2.2.9) then the Runge-Kutta defining equations (3.2.1-2) are soluble for any $y_n \in \mathbb{R}^m$ and any $h > 0$.

Proof. By Result 3.6.10 (ii) A is invertible and $\Psi_0(A^{-1}) > 0$. Now by the definition of $\Psi_0(A^{-1})$ there exists a positive definite diagonal matrix D such that

$$\Psi_D(A^{-1}) > 0$$

and by scaling we can choose D such that $\sum_{i=1}^s d_i = 1$. Let $\Phi(Y)$ be defined by (3.6.13) and consider (3.6.14) once again. Bounds on the first two terms of the right-hand side of (3.6.14) are given by (3.6.15) and (3.6.16). For the third term note that

$$\begin{aligned} Y^T(D \otimes I_m)F(Y) &= \langle Y, F(Y) \rangle_D \\ &= \sum_{i=1}^s d_i \langle Y_i, f(Y_i) \rangle \end{aligned}$$

and hence by (2.2.9)

$$Y^T(D \otimes I_m)F(Y) \leq w \quad (3.6.26)$$

where

$$w = \sup_{\|x\| \leq R} \langle x, f(x) \rangle$$

which is finite since f is continuous and the supremum is taken over a compact set. Hence we have that

$$\langle Y, \Phi(Y) \rangle_D > \Psi_D(A^{-1}) \|Y\|_D^2 - \|Y\|_D \|(A^{-1} \otimes I_m)y\|_D - w$$

and thus $\langle Y, \Phi(Y) \rangle_D > 0$ for $\|Y\|_D$ sufficiently large, and in particular $\langle Y, \Phi(Y) \rangle_D > 0$ for all $Y \in \partial B$ where B is the ball of radius r , for r sufficiently large, in $(\mathbb{R}^{ms}, \|\cdot\|_D)$ centred at the origin. Thus this ball is forward invariant for the dynamical system defined by

$$\frac{dY}{dt} = -\Phi(Y)$$

and it follows from Result 2.6.2 that there exists $Y \in B$ such that $\Phi(Y) = 0$ which defines the required solution of (3.2.1). ■

Remark (i) Note that (3.6.25) implies that A is invertible.

(ii) Since (2.2.9) is a generalization of (2.2.1), Example 3.6.19 implies that the solution of the Runge-Kutta defining equations (3.2.1–2) is not in general unique.

(iii) There exist methods of arbitrarily high order that satisfy (3.6.25), including in particular the Butcher IA, Radau IA and Radau IIA classes of quadrature based methods.

We now show that there always exists a solution sequence $\{y_n\}_{n=0}^\infty$ when (2.1.1, 2.2.9) is approximated numerically using either the one or two stage theta method.

Corollary 3.6.21 *If f satisfies (2.2.9) then the equations defining the one-stage theta method (3.5.1) and the two-stage theta method (3.5.2) are soluble for any $\mathbf{y}_n \in \mathbb{R}^m$, any $h > 0$ and any $\theta \in [0, 1]$.*

Proof. First consider the one-stage theta method. The result is trivial if $\theta = 0$ since the method is then explicit. For $\theta \in (0, 1]$ $A = \theta$ and it follows easily from (3.6.1–2) that $\Psi_0(A) = \theta$, and hence by Proposition 3.6.20 the defining equations are soluble.

The result now also follows for the two-stage theta method from the equivalence of solution sequences to these two methods given in Example 3.6.6 or by noting, as in the proof of Corollary 3.6.3 that the solution of the defining equations for the two-stage theta method is equivalent to solving the defining equation for the one-stage theta method with a perturbed \mathbf{y}_n . ■

3.7 Linear Multistep Methods as Dynamical Systems

Linear multistep methods do not naturally define discrete dynamical systems on \mathbb{R}^m in the same way as Runge-Kutta methods. This is because these methods use previous values of \mathbf{y}_n . If (3.3.1) is uniquely soluble then the linear multistep method does define a discrete dynamical system on \mathbb{R}^{mk} , where (3.3.1) defines a map from \mathbb{R}^{mk} to \mathbb{R}^{mk} by

$$[\mathbf{y}_n, \mathbf{y}_{n+1}, \dots, \mathbf{y}_{n+k-1}]^T \mapsto [\mathbf{y}_{n+1}, \mathbf{y}_{n+2}, \dots, \mathbf{y}_{n+k}]^T.$$

To compare the dynamics of the numerical solution with that of the underlying problem, following this approach it is necessary to compare a dynamical system on \mathbb{R}^m with a discrete dynamical system on \mathbb{R}^{mk} . This adds an extra complication to the case for Runge-Kutta methods where the two dynamical systems were defined on the same space, and we will not adopt this approach for linear multistep methods.

Suprisingly Kirchgraber [41] showed that every strongly stable linear multistep method is equivalent, in some sense, to some one-step method, and so these methods can after all define dynamical systems on \mathbb{R}^m . However Kirchgraber's proof is implicit and so it is hard to make use of this result.

More recently Eirola and Nevanlinna [15] considered linear multistep methods directly as mappings from \mathbb{R}^{mk} to \mathbb{R}^m and showed that as such they are closely related to a map which is expressible directly in terms of the underlying flow.

3.7.1 Solubility of Linear Multistep Defining Equations

If $\beta_k \neq 0$ then the linear multistep method (3.3.1) is implicit, and we need to determine whether and under what conditions the implicit equations defining the method are soluble. Such results though, follow directly from the existence and uniqueness results for Runge-Kutta methods, as we now show. Consider (3.3.1) and let

$$\tilde{y}_{n+k-1} = h \sum_{j=0}^{k-1} \beta_j f(y_{n+j}) - \sum_{j=0}^{k-1} \alpha_j y_{n+j}$$

represent the known values in (3.3.1) and define $\tilde{h} = h\beta_k$. Then (3.3.1) can be rewritten as

$$y_{n+k} = \tilde{y}_{n+k-1} + \tilde{h}f(y_{n+k}). \quad (3.7.27)$$

But (3.3.1) is simply the backward Euler method, and hence solubility theory for implicit linear multistep methods follows directly from the theory that we have already derived for implicit Runge-Kutta methods.

Chapter 4

Spurious Limit Sets

4.1 Introduction

In this chapter we consider the existence and effect of so-called spurious limit sets. The asymptotic behaviour of a dynamical system is given by its ω -limit sets. If the limit sets of the underlying system and its numerical approximation are different, then clearly so will be the dynamics of the two systems, and thus for a numerical method to reproduce the correct asymptotic behaviour it is essential that the ω -limit sets of the numerical approximation are “close” to the corresponding ω -limit sets of the underlying system.

In this chapter we consider the solution of the autonomous initial value problem: find $\mathbf{y} \in \mathbb{R}^m$ satisfying

$$\frac{d\mathbf{y}}{dt} = \mathbf{f}(\mathbf{y}) \quad \text{for } t \geq 0 \quad \text{and} \quad \mathbf{y}(0) = \mathbf{y}_0 \quad (4.1.1)$$

where $\mathbf{f}: \mathbb{R}^m \rightarrow \mathbb{R}^m$, although we also consider nonautonomous problems in Section 4.5. Continuity conditions on \mathbf{f} will be stated where required.

The simplest ω -limit sets of (4.1.1) are fixed points (also called steady solutions). Iserles [38] showed that Runge-Kutta and linear multistep methods retain all the fixed points of (4.1.1), however some Runge-Kutta methods (but not linear multistep methods) may generate additional fixed points which do not correspond to fixed points of (4.1.1). These additional steady solutions introduced by the discretization are referred to as *spurious fixed points*.

Some Runge-Kutta and linear multistep methods also admit solutions of the form $\mathbf{y}_{2n} = \mathbf{u}$, $\mathbf{y}_{2n+1} = \mathbf{v} \forall n \geq 0$, where $\mathbf{u} \neq \mathbf{v}$. This is known as a *period two solution*

(or *two-cycle* or *sawtooth solution*). Such periodic motion on the grid scale must be spurious.

If numerical discretization admits spurious fixed points or two-cycles then the ω -limit sets of the underlying system and the numerical approximation will not correspond and, at least for certain initial conditions, the numerical solution will display incorrect asymptotic behaviour. If the spurious solutions are stable then they may attract a large set of initial conditions, and in such a case the numerical approximation ceases to be an “approximation” to the underlying system over long time intervals.

Whilst two-cycles are easy to recognise as spurious, it should be noted that solutions converging to spurious fixed points are often smooth, and may not at first sight appear spurious. The unwary may mistake such solutions for genuine solutions of the underlying system.

Although an unstable spurious solution will not attract a large set of initial data, such solutions are also undesirable. This is because, as has been noted in Stuart [53], the unstable manifold of the spurious solution is often connected to infinity, and thus the existence of an unstable spurious fixed point or two-cycle may cause the numerical solution to blow up; see also Elliott and Stuart [17]. If this happens then the structure of the underlying system will be lost. For example the discrete system defined by the numerical approximation of a dissipative system will not be dissipative and will not possess a global attractor if it has a spurious solution whose unstable manifold is connected to infinity.

The observations above led to an attempt to classify the methods which do not admit spurious fixed points and/or two-cycles. The following definitions are reproduced from [39].

Definition 4.1.1 A numerical method for (4.1.1) which does not admit spurious fixed points is said to be *regular of degree 1*, denoted $R^{[1]}$. A method which is not $R^{[1]}$ is said to be *irregular of degree 1*, denoted $IR^{[1]}$.

Definition 4.1.2 A numerical method for (4.1.1) which does not admit period two solutions is said to be *regular of degree 2*, denoted $R^{[2]}$. A method which is not $R^{[2]}$ is said to be *irregular of degree 2*, denoted $IR^{[2]}$.

We will also use the notation $R^{[1,2]}$ to denote a method which is $R^{[1]}$ and $R^{[2]}$, etc. Examples of spurious fixed point and period two solutions, and their effect on the

dynamics of the numerical map can be found in [23, 38, 39, 46, 53, 57]. These spurious solutions often bifurcate from the linear stability limit, but it should be noted that they can persist for arbitrarily small values of the step-size h , and thus incorrect asymptotic behaviour can be observed at step-sizes used in practical implementations.

A thorough study of regular Runge-Kutta and linear multistep methods has been conducted in [25, 38, 39, 40, 57]. Iserles [38] presents examples of spurious steady solutions of Runge-Kutta, linear multistep and predictor-corrector methods arising from Riccati equations, and also shows that all linear multistep methods are $R^{[1]}$. Stuart and Peplow [57] classify the $R^{[1,2]}$ two-stage theta methods (3.5.2), and study the period two solutions of $IR^{[2]}$ methods. That paper was also the first to consider the existence of spurious solutions of irregular methods in the limit as $h \rightarrow 0$. It was shown that if $f \in C^1(\mathbb{R}^m, \mathbb{R}^m)$, then period two solutions of $IR^{[2]}$ two-stage theta methods become unbounded as $h \rightarrow 0$, if they exist for h arbitrarily small. This result, which is a special case of Theorem B(c)(ii) below, inspires the approach of this chapter. Hairer, Iserles and Sanz-Serna [25] conduct a systematic study of the spurious equilibria of Runge-Kutta methods, and in particular classify all the $R^{[1]}$ Runge-Kutta methods by means of a recursive test. Iserles, Peplow and Stuart [39] present a unified theory of spurious solutions based on local bifurcation theory, using the step-size h as the bifurcation parameter. Amongst many other results they show that the maximum order of a $R^{[1,2]}$ Runge-Kutta method is 2, and that the recursive test of [25] can be used to classify these methods. Also in that paper all the $R^{[2]}$ linear multistep methods are identified and the regularity properties of a class of predictor-corrector methods are studied. In [40] Iserles and Stuart consider $R^{[1,2]}$ linear multistep methods further, and a modification of the backward differentiation formulae which generates such methods is proposed.

Other considerations mean that (4.1.1) is often numerically integrated using a method which is not $R^{[1,2]}$. For example the highest possible order of a $R^{[1,2]}$ Runge-Kutta method is 2, and Hairer *et al* [25] proved that the forward Euler method is the only $R^{[1]}$ explicit Runge-Kutta method. If a method which is not $R^{[1,2]}$ is used, then spurious solutions may exist, and to ensure good numerical reproduction of the dynamics of (4.1.1) it is necessary to study the existence of spurious solutions in irregular methods. It is this approach, complimentary to the study of regular methods *per se*, which we will follow in this chapter.

Although we only consider fixed time-stepping methods, we will treat the step-size h as a bifurcation or continuation parameter and consider the existence of spurious fixed point and period two solutions in the limit as $h \rightarrow 0$. Simple continuity conditions will be applied to \mathbf{f} in (4.1.1), which will allow us to derive results on the possible existence and boundedness of spurious solutions in the limit as $h \rightarrow 0$. The main results are stated below.

Theorem A *If (4.1.1) is approximated numerically using a Runge-Kutta method, where for an implicit method the solution of (3.2.1) constructed in Proposition 3.6.3 is used, then*

- (i) *if \mathbf{f} is globally Lipschitz there exists $h_c > 0$ such that if $h \in (0, h_c)$ the numerical solution does not admit any spurious fixed points,*
- (ii) *if \mathbf{f} is locally Lipschitz, and spurious fixed points exist for h arbitrarily small, then these spurious fixed points tend to infinity, in norm, as $h \rightarrow 0$,*
- (iii) *if $\mathbf{f} \in C(\mathbb{R}^n, \mathbb{R}^n)$ and a continuous branch $\tilde{\mathbf{y}}(h)$ of fixed point solutions of the numerical method exists for h sufficiently small, then as $h \rightarrow 0$, either $\|\tilde{\mathbf{y}}(h)\| \rightarrow \infty$ or $\|\mathbf{f}(\tilde{\mathbf{y}}(h))\| \rightarrow 0$. If furthermore the zeros of \mathbf{f} are isolated then $\|\mathbf{f}(\tilde{\mathbf{y}}(h))\| \rightarrow 0$ implies $\tilde{\mathbf{y}}(h) \rightarrow \tilde{\mathbf{y}}$, a fixed point of (4.1.1),*
- (iv) *if a spurious fixed point solution bifurcates from $\tilde{\mathbf{y}}$ at $h = 0$ then either*
 - \mathbf{f} is not continuous at $\tilde{\mathbf{y}}$, or,*
 - $\mathbf{f}(\tilde{\mathbf{y}}) = 0$ and \mathbf{f} is not Lipschitz at $\tilde{\mathbf{y}}$.*

Theorem B *If (4.1.1) is approximated numerically using a Runge-Kutta method, where for an implicit method the solution of (3.2.1) constructed in Proposition 3.6.3 is used, or a zero-stable linear multistep method of the form (4.1.1), with $\rho(-1) \neq 0$, then*

- (i) *if \mathbf{f} is globally Lipschitz there exists $h_c > 0$ such that if $h \in (0, h_c)$ the numerical solution does not admit any period two solutions,*
- (ii) *if \mathbf{f} is locally Lipschitz and a period two solution $(\mathbf{u}(h), \mathbf{v}(h))$ exists for h arbitrarily small, then $\mathbf{u}(h), \mathbf{v}(h)$ both tend to infinity, in norm, as $h \rightarrow 0$,*
- (iii) *if $\mathbf{f} \in C(\mathbb{R}^n, \mathbb{R}^n)$ and a continuous branch $(\mathbf{u}(h), \mathbf{v}(h))$ of period two solutions of the numerical method exists for h arbitrarily small, then as $h \rightarrow 0$, $\|\mathbf{u}(h)\|, \|\mathbf{v}(h)\|$ both tend to infinity, or $\|\mathbf{f}(\mathbf{u}(h))\|, \|\mathbf{f}(\mathbf{v}(h))\|$ and $\|\mathbf{u}(h) - \mathbf{v}(h)\| \rightarrow 0$. If furthermore the zeros of \mathbf{f} are isolated then $\|\mathbf{f}(\mathbf{u}(h))\| \rightarrow 0$ implies $\mathbf{u}(h), \mathbf{v}(h)$ tend to $\tilde{\mathbf{y}}$, a fixed point of (4.1.1),*

(iv) if a period two solution bifurcates from \tilde{y} at $h = 0$ then either

f is not continuous at \tilde{y} , or,

$f(\tilde{y}) = 0$ and f is not Lipschitz at \tilde{y} .

The proofs of the above results can be found in the following sections, where sufficient bounds on the step-size h to prevent spurious solutions in the case where f is globally Lipschitz are also given, and many other results can also be found.

In Section 4.2 we develop the theory for spurious fixed point solutions of Runge-Kutta methods. In addition to the results above, several corollaries are also given. Example 4.2.4 shows that Theorems A and B do not apply to arbitrary solutions of implicit Runge-Kutta methods. This implies that some assumption on the solution of the Runge-Kutta defining equations, such as is made above, is necessary for implicit Runge-Kutta methods.

The theory is extended to cover period two solutions of Runge-Kutta and linear multistep methods in Sections 4.3 and 4.4. The Runge-Kutta results follow easily from those in Section 4.2, whilst the linear multistep results follow from Lemma 4.4.3 which shows that for fixed step-size there is at most one two-cycle of any linear multistep method passing through any point of \mathbb{R}^m . In Example 4.3.5 a continuous initial value problem that generates bounded spurious solutions for h arbitrarily small is presented, showing that Theorems A(iii) and B(iii) are relevant.

In Section 4.5 we will consider spurious solutions of nonautonomous systems. We will show that methods which are regular for autonomous systems can admit spurious solutions when applied to nonautonomous problems. Specifically we will present an example where the trapezoidal method, which is $R^{[1,2]}$, admits a spurious fixed point, and another where it admits a two-cycle. We will also extend Theorem A (i) and (ii) to cover spurious fixed points of Runge-Kutta methods applied to a certain class of nonautonomous problems, and in so doing will illustrate the issues involved in extending our results to nonautonomous systems.

Theorems A and B suggest that when f satisfies a Lipschitz condition on a bounded set B , spurious solutions will not degrade the numerical solution on B for h sufficiently small (depending on B). However this will not be true on the whole of \mathbb{R}^m since, as noted above, an unstable spurious solution can destroy a global attractor, and even though the spurious solution becomes unbounded as $h \rightarrow 0$ the dynamics of the continuous and numerical systems will differ significantly for some initial conditions however

small the step-size is.

By Theorem A(iv) and B(iv) even for arbitrarily small step-sizes we cannot be sure that a numerical method will produce the correct behaviour in a neighbourhood of a fixed point where f is not Lipschitz. However it should be noted that the solution of (4.1.1) itself is not unique in a neighbourhood of such a point.

In seeking to prove general results, no assumption has been made at any stage on the global structure of the nonlinear function f , and hence our results apply to all problems of the form (4.1.1). It should be noted then, that in some cases and for some methods it can be shown that spurious solutions cannot exist for h arbitrarily small, although f is not globally Lipschitz, but where some other structure is imposed on the nonlinear term. For example, we will see in Chapter 6 that when a dissipative problem of the form (6.1.1–2) is solved numerically using a Runge-Kutta method then it is sufficient for f to be locally Lipschitz to imply that that spurious fixed points cannot exist for h arbitrarily small.

In Chapter 7 we will extend Theorems A(iii) and B(iii) to cover arbitrary invariant sets, not just fixed points and two-cycles. Theorem 7.4.7 shows that if numerically invariant sets converge to a compact set as $h \rightarrow 0$ then this set is invariant under the evolution of (4.1.1), and thus numerical invariant sets either converge to an invariant set of the underlying system or become unbounded as $h \rightarrow 0$.

4.2 Spurious Fixed Points of Runge-Kutta Methods

In this section the spurious fixed point solutions of explicit and implicit Runge-Kutta methods are considered. We will often assume a Lipschitz continuity condition on f , but a series of ε, δ -arguments will enable us to prove some results when Lipschitz conditions do not apply. The $R^{[1]}$ methods were classified in [25] by a recursive test. A simple classification for explicit methods was found:

Result 4.2.1 (Hairer, Iserles & Sanz-Serna [25]) *A consistent explicit Runge-Kutta method of the form (3.2.1–2) is $R^{[1]}$ if and only if it produces the same solution sequence as the forward Euler method (3.1.1). ■*

The solution to (4.1.1) is often approximated using a high order explicit Runge-Kutta method. By Theorem 4.2.1, such a method is necessarily $IR^{[1]}$, and we may expect spurious steady solutions, and hence incorrect dynamics. This motivates our

approach of considering the spurious solutions of irregular methods, rather than simply classifying the regular methods. Note that if $\mathbf{A} = 0$ (where \mathbf{A} is defined by (3.2.7)) then $a_{ij} = 0$ for all i, j and the method produces the same solution sequence as the forward Euler method, and hence is $R^{[1]}$, and admits no spurious fixed points. We now prove that if $\mathbf{A} \neq 0$ then spurious fixed points cannot exist for h arbitrarily small if \mathbf{f} is globally Lipschitz.

Theorem 4.2.2 *If \mathbf{f} is globally Lipschitz, with Lipschitz constant L , and*

$$h < \frac{1}{L\mathbf{A}(1 + \mathbf{B})}, \quad (4.2.1)$$

where \mathbf{A} and \mathbf{B} are defined by (3.2.7) and (3.2.8) and $\mathbf{A} > 0$, then the Runge-Kutta method (3.2.1–2) admits no spurious fixed points when applied to (4.1.1).

Proof. Suppose there exists a solution of (3.2.1–2) such that $\mathbf{y}_n = \mathbf{y}_{n+1}$ with $\mathbf{f}(\mathbf{y}_n) \neq 0$. Since \mathbf{f} is globally Lipschitz, Lemma 3.2.3 applies and (3.2.16) holds. Now (3.2.2) implies

$$\sum_{i=1}^s b_i \mathbf{f}(\mathbf{Y}_i) = 0 \quad (4.2.2)$$

and by consistency

$$\begin{aligned} \|\mathbf{f}(\mathbf{y}_n)\| &= \left\| \sum_{i=1}^s b_i \mathbf{f}(\mathbf{y}_n) \right\| \\ &= \left\| \sum_{i=1}^s b_i (\mathbf{f}(\mathbf{y}_n) - \mathbf{f}(\mathbf{Y}_i)) \right\| \\ &\leq \mathbf{B} \max_i \|\mathbf{f}(\mathbf{y}_n) - \mathbf{f}(\mathbf{Y}_i)\| \\ &\leq \frac{Lh\mathbf{A}\mathbf{B}}{1 - hL\mathbf{A}} \|\mathbf{f}(\mathbf{y}_n)\| \quad \text{by (3.2.16)} \end{aligned} \quad (4.2.3)$$

and since (4.2.1) holds, (4.2.3) implies $\|\mathbf{f}(\mathbf{y}_n)\| < \|\mathbf{f}(\mathbf{y}_n)\|$, clearly a contradiction. ■

There are very few interesting problems of the form (4.1.1) for which \mathbf{f} is globally Lipschitz, but \mathbf{f} is often locally Lipschitz, and we would like to generalise Theorem 4.2.2 to this case. The following example modified from an example in [53] shows that this cannot be done.

Example 4.2.3 Consider the initial value problem

$$\frac{dy}{dt} = -y^3, \quad \text{where } y(0) \in \mathbf{R}. \quad (4.2.4)$$

The origin is the only fixed point of (4.2.4). Now suppose a numerical approximation is obtained using the forward Euler method (3.1.1). This implies that

$$y_{n+1} = y_n - hy_n^3. \quad (4.2.5)$$

It is simple to check that $y_n = (-1)^n \sqrt{2/h}$ defines a period two solution of (4.2.5). Now suppose that the numerical solution is obtained using the method (4.2.6)

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ \hline & 1/2 & 1/2 \end{array} \quad (4.2.6)$$

One step of this method with step-size h corresponds to two steps of the forward Euler method with step-size $h/2$. Thus for any $h > 0$ $y_n = \sqrt{4/h}$ and $y_n = -\sqrt{4/h}$ are both spurious fixed points of the method (4.2.6) for the problem (4.2.4). \square

Notice that the spurious solutions in the example exist for all h , but tend to infinity as $h \rightarrow 0$. We will later prove that if f is locally Lipschitz then two-cycles of linear multistep methods which exist for arbitrarily small h tend to infinity as $h \rightarrow 0$. The spurious fixed points in the example correspond to period two solutions of the forward Euler method, and hence must tend to infinity as $h \rightarrow 0$. We would also like to prove this result for general spurious fixed points of Runge-Kutta methods, but the example below shows that it does not hold without further assumptions.

Example 4.2.4 Consider the initial value problem

$$\frac{dy}{dt} = f_1(y) = y^3, \quad \text{where } y(0) \in \mathbb{R}. \quad (4.2.7)$$

The origin is the only fixed point of (4.2.7). Now suppose a numerical approximation is obtained using the Runge-Kutta method (4.2.8).

$$\begin{array}{c|cc} 1 & 0.75 & 0.25 \\ 1 & 0.25 & 0.75 \\ \hline & 0.5 & 0.5 \end{array} \quad (4.2.8)$$

Let $y_n = 0$, then it is simple to check that (4.2.9) solves the equations (3.2.1–2) for the method (4.2.8) and any $h > 0$ where \mathbf{f} is given by f_1 .

$$\left. \begin{aligned} y_n &= y_{n+1} = 0 \\ Y_1 &= \sqrt{2/h} \\ Y_2 &= -\sqrt{2/h} \end{aligned} \right\} \quad (4.2.9)$$

Now consider the modified problem

$$\frac{dy}{dt} = f_2(y) = (y + p(y))^3, \quad \text{where } y(0) \in \mathbb{R} \quad (4.2.10)$$

and p is the test function

$$p(y) = \begin{cases} 0 & \text{if } |y| \geq 1 \\ \exp\left[\frac{1}{y^2-1}\right] & \text{if } |y| < 1. \end{cases}$$

Observe that $f_2(0) = e^{-3} \neq 0$, thus the origin is not a fixed point of (4.2.10), also since p is a test function, (see [21]), $f_2 \in C^\infty(\mathbb{R}, \mathbb{R})$. Now consider the numerical solution using the Runge-Kutta method (4.2.8). For $|y| \geq 1$, $f_1(y) = f_2(y)$, therefore if $h \leq 2$ and $y_n = 0$ then (4.2.9) also solves the equations (3.2.1–2) for the modified problem (4.2.10). Thus we have a problem of the form (4.1.1), where \mathbf{f} is smooth, and a Runge-Kutta method which generates a spurious fixed point which exists for h arbitrarily small and is itself fixed as $h \rightarrow 0$. \square

All hope is not lost however. It should be noted that in both the problems considered in Example 4.2.4 the equations (3.2.1) admit more than one solution. In Proposition 3.6.3 we proved that if \mathbf{f} is locally Lipschitz then there is a locally unique solution of (3.2.1) in a neighbourhood of \mathbf{y}_n . For f_1 this locally unique solution is $y_n = y_{n+1} = Y_1 = Y_2 = 0$ which is far more ‘natural’ than the solution given in the example. We also proved in Proposition 3.6.3 that the iteration scheme (3.6.2) converges to the natural solution of (3.2.1), and hence if (3.6.2) is used to solve the Runge-Kutta defining equations (3.2.1) then the spurious fixed point seen in Example 4.2.4 will not arise. Where the implicit equations are not uniquely soluble, we will need to assume that the solution of (3.2.1) defined by Proposition 3.6.3 is used when \mathbf{f} is locally Lipschitz, and a solution of (3.2.1) defined by Proposition 3.6.2 is used when \mathbf{f} is continuous, to enable us to prove results about the existence of spurious solutions

in these cases. This assumption will be explicitly stated where it is made. In addition to the iteration scheme (3.6.2) we claim that any ‘sensible’ iteration scheme used in practical implementations will converge to the ‘natural’ solution of the equations, and hence that our results will apply.

In this chapter we will not make any structural assumptions on \mathbf{f} . Under certain structural assumptions on \mathbf{f} it is possible to prove that the Runge-Kutta defining equations are uniquely soluble and hence that the assumptions mentioned above are not always necessary. Of course, for an explicit method (3.2.1) is trivially uniquely soluble, and none of these assumptions are needed for explicit methods.

It should be noted that Propositions 3.6.2 and 3.6.3 ensure the existence of a solution of (3.2.1) for one step if h is sufficiently small, but do not guarantee that a solution sequence $\{\mathbf{y}_n\}_{n=0}^{\infty}$ can be generated with fixed step-size. But, since our aim in this chapter is to study the existence of spurious solutions, we will assume that a solution sequence exists, with the equations (3.2.1–2) being solved exactly, then Proposition 3.6.2 and Proposition 3.6.3 will enable us to derive results on the nature of the spurious solutions.

All the remaining results in this section will follow from the two lemmas below.

Lemma 4.2.5 *If $B \subset \mathbb{R}^m$ is bounded and \mathbf{f} is Lipschitz on $\mathcal{N}(B, \delta)$ with Lipschitz constant L for some $\delta > 0$, $A > 0$ and*

$$h < \min \left(\frac{\delta}{aM}, \frac{1}{LA(1 + \mathbb{B})} \right) \quad (4.2.11)$$

where

$$M = \sup_{\mathbf{y} \in \mathcal{N}(B, \delta)} \|\mathbf{f}(\mathbf{y})\| \quad (4.2.12)$$

then for $\mathbf{y}_n \in B$ the solution of the Runge-Kutta method (3.2.1–2) satisfies $\mathbf{y}_n = \mathbf{y}_{n+1}$ if and only if $\mathbf{f}(\mathbf{y}_n) = 0$, where for an implicit method we assume that the solution of (3.2.1) defined by Proposition 3.6.3 is used.

Proof. Proposition 3.6.3 implies that the conditions for Lemma 3.2.3 hold and hence (3.2.16) holds. Now suppose that there exists a solution of (3.2.1–2) such that $\mathbf{y}_n \in B$, $\mathbf{y}_n = \mathbf{y}_{n+1}$ and $\mathbf{f}(\mathbf{y}_n) \neq 0$ then follow the proof of Theorem 4.2.2 from (4.2.2) to obtain the result. ■

Lemma 4.2.6 *If $B \subset \mathbb{R}^m$ is bounded, f is continuous on $\overline{\mathcal{N}}(B, \delta)$ for some $\delta > 0$ and if $\mathbb{A} > 0$ then given any $\varepsilon > 0$ there exists $H(\varepsilon) > 0$ such that for $h < H(\varepsilon)$ any fixed point solution $\hat{\mathbf{y}}$ of (3.2.1–2) with $\hat{\mathbf{y}} \in B$ satisfies $\|f(\hat{\mathbf{y}})\| < \varepsilon$, where for an implicit method we assume that the solution of (3.2.1) defined by Proposition 3.6.3 is used.*

Proof. Since $\overline{\mathcal{N}}(B, \delta)$ is compact, f is uniformly continuous on $\overline{\mathcal{N}}(B, \delta)$, so given $\hat{\varepsilon} > 0$ $\exists \hat{\delta} > 0$ such that any $\mathbf{x}, \mathbf{y} \in \overline{\mathcal{N}}(B, \delta)$ with $\|\mathbf{x} - \mathbf{y}\| < \hat{\delta}$ satisfy $\|f(\mathbf{x}) - f(\mathbf{y})\| < \hat{\varepsilon}$. Let $\hat{\varepsilon} = \varepsilon/\mathbb{B}$ and $\tilde{\delta} = \min(\hat{\delta}, \delta)$. By Proposition 3.6.2 if $h < \tilde{\delta}/aM$ where M is defined by (4.2.12) then for any $\mathbf{y}_n \in B$ it follows that $\|\mathbf{Y}_i - \mathbf{y}_n\| < \tilde{\delta}$. Hence if $\mathbf{y}_n \in B$ is a fixed point solution of (3.2.1–2) it follows that;

$$\begin{aligned} \|f(\mathbf{y}_n)\| &= \left\| \sum_{i=1}^s b_i f(\mathbf{y}_n) \right\| \\ &= \left\| \sum_{i=1}^s b_i (f(\mathbf{y}_n) - f(\mathbf{Y}_i)) \right\| \\ &< \mathbb{B} \hat{\varepsilon} \\ &= \varepsilon \end{aligned}$$

as required. ■

Example 4.2.3 showed that it is possible for spurious solutions to exist for h arbitrarily small when f is locally Lipschitz, and we can now prove that such spurious solutions tend to infinity as $h \rightarrow 0$.

Theorem 4.2.7 *If f is locally Lipschitz and spurious fixed point solutions of (3.2.1–2) exist for h arbitrarily small then these tend to infinity in norm as $h \rightarrow 0$, where for an implicit method we use the solution of (3.2.1) defined by Proposition 3.6.3. By this we mean that if there exists a sequence (\mathbf{u}_p, h_p) such that $h_p > 0 \forall p$, $h_p \rightarrow 0$ as $p \rightarrow \infty$ and \mathbf{u}_p is a spurious fixed point of the method with step-size h_p then $\|\mathbf{u}_p\| \rightarrow \infty$ as $p \rightarrow \infty$.*

Proof. It is sufficient to prove that for any bounded set B , for h sufficiently small no point of B is a spurious fixed point, but this follows trivially from Lemma 4.2.5. ■

An alternative statement of Theorem 4.2.7 is to say that if bounded sequence or continuous branch of fixed point solutions exists as $h \rightarrow 0$ then $\exists H > 0$ such that for $h < H$ the corresponding fixed point solution of (3.2.1–2), \mathbf{u} , satisfies $f(\mathbf{u}) = 0$, that is \mathbf{u} is a fixed point solution of (4.1.1). If we relax the condition that that f is

Lipschitz continuous and assume merely that \mathbf{f} is continuous on \mathbb{R}^m then the following theorem shows that continuous branches of spurious fixed point solutions which exist for h arbitrarily small, either tend to steady solutions of the underlying differential equation, or diverge to infinity as $h \rightarrow 0$.

Theorem 4.2.8 *Suppose $\mathbf{f} \in C(\mathbb{R}^m, \mathbb{R}^m)$ and there exists a continuous branch of fixed points $\mathbf{u}(h)$ of (3.2.1–2) for $h \in (0, H]$, where for an implicit method the solution of (3.2.1) is defined by Proposition 3.6.2, then as $h \rightarrow 0$ either*

- (i) $\|\mathbf{u}(h)\| \rightarrow \infty$, or,
- (ii) $\|\mathbf{f}(\mathbf{u}(h))\| \rightarrow 0$.

If furthermore the zeros of \mathbf{f} are isolated then as $h \rightarrow 0$ either (i) occurs or $\mathbf{u}(h) \rightarrow \tilde{\mathbf{y}}$ as $h \rightarrow 0$, where $\tilde{\mathbf{y}}$ is a fixed point of (4.1.1).

Proof. To show that either (i) or (ii) occurs it is sufficient to show that as $h \rightarrow 0$, for any bounded set B , $\|\mathbf{f}(\mathbf{u}(h))\| \rightarrow 0$ or for all sufficiently small h , $\mathbf{u}(h)$ is not in B , but this follows trivially from Lemma 4.2.6. If the zeros of \mathbf{f} are isolated and $\mathbf{u}(h)$ remains bounded as $h \rightarrow 0$ then the last part follows by the continuity of \mathbf{f} . ■

Remark Example 4.3.5 shows that if \mathbf{f} is continuous on \mathbb{R}^m it is possible for a Runge-Kutta method to generate a spurious fixed point solution which remains bounded and which converges to a steady solution of (4.1.1) as $h \rightarrow 0$.

The following theorem gives a necessary condition for the bifurcation of spurious fixed point solutions from $\tilde{\mathbf{y}}$ at $h = 0$, namely either

- (a) \mathbf{f} is not continuous at $\tilde{\mathbf{y}}$, or,
- (b) $\mathbf{f}(\tilde{\mathbf{y}}) = 0$ and \mathbf{f} is not Lipschitz at $\tilde{\mathbf{y}}$.

Theorem 4.2.9 *Suppose there exists a sequence (\mathbf{u}_p, h_p) such that $h_p > 0 \forall p$, $h_p \rightarrow 0$, and $\mathbf{u}_p \rightarrow \tilde{\mathbf{y}}$ as $p \rightarrow \infty$ where, for each p , \mathbf{u}_p is a spurious fixed point solution of (3.2.1–2) with step-size h_p , and if the method is implicit then the solution of (3.2.1) is defined by Proposition 3.6.2 then if \mathbf{f} is continuous on a neighbourhood of $\tilde{\mathbf{y}}$ it follows that*

- (i) $\mathbf{f}(\tilde{\mathbf{y}}) = 0$, that is $\tilde{\mathbf{y}}$ is a steady solution of (4.1.1), and,
- (ii) \mathbf{f} is not Lipschitz at $\tilde{\mathbf{y}}$.

Proof. (i) By Lemma 4.2.6 $\|f(u_p)\| \rightarrow 0$ as $p \rightarrow \infty$, and result follows by continuity of f .

(ii) Follows trivially from Lemma 4.2.5. ■

Remark The above theorem also shows that if the numerical solution is asymptotic to \tilde{y} for arbitrarily small h then \tilde{y} is a genuine asymptotic fixed point of (4.1.1), (although it does not necessarily follow that the solution of the continuous problem is asymptotic to \tilde{y} if the same initial value is used as for the numerical method).

Now consider general f but suppose f is Lipschitz on $U \subset \mathbb{R}^m$. For given $\delta > 0$ define $B \subset U$ by

$$B = \{x \in U : \text{dist}(x, U^c) \geq \delta\} \quad (4.2.13)$$

where U^c is the complement of U in \mathbb{R}^m . This implies that $\mathcal{N}(B, \delta) \subseteq U$. By Lemma 4.2.5 if $\hat{y} \in U$ is a spurious fixed point solution of (3.2.1–2) and $h < H(\delta)$ then $\hat{y} \notin B$ and hence $\text{dist}(\hat{y}, U^c) < \delta$, so that is \hat{y} is within distance δ of the boundary U . We can force spurious fixed points to the boundary of U by taking δ as small as we like. By (3.6.4) as $\delta \rightarrow 0$, $H(\delta) \rightarrow 0$. In this way we prove that as $h \rightarrow 0$ spurious fixed points either ‘converge’ to the set on which f is not Lipschitz or ‘diverge’ to infinity.

Corollary 4.2.10 *Suppose f is Lipschitz on every bounded subset of some set D , and that for an implicit method the solution of (3.2.1) defined by Proposition 3.6.3 is used, then given any positive δ, β there exists $H(\delta, \beta) > 0$ such that every spurious fixed point \hat{y} of (3.2.1–2) with $h < H(\delta, \beta)$, satisfies either*

- i) $\text{dist}(\hat{y}, D^c) < \delta$, or,
- ii) $\|\hat{y}\| > \beta$.

Proof. Apply Lemma 4.2.5 with $U = \overline{B}(0, \beta + \delta) \cap D$ and B defined by (4.2.13). ■

If the Lipschitz continuity condition is dropped the following result holds.

Corollary 4.2.11 *Suppose f is continuous on every bounded subset of some set D , and that for an implicit method a solution of (3.2.1) defined by Proposition 3.6.2 is used, then given any positive $\varepsilon, \beta, \delta$ there exists $H(\varepsilon, \beta, \delta) > 0$ such that every spurious fixed point \hat{y} of (3.2.1–2) with $h < H(\varepsilon, \beta, \delta)$, satisfies either*

- i) $\text{dist}(\hat{y}, D^c) < \delta$, or,

- ii) $\|\hat{\mathbf{y}}\| > \beta$, or,
- iii) $\|f(\hat{\mathbf{y}})\| < \varepsilon$,

Proof. Apply Lemma 4.2.6 with U and B defined as in the proof of Corollary 4.2.10. ■

4.3 Spurious Two-Cycles of Runge-Kutta methods

In this section we will derive results for (spurious) two-cycles of Runge-Kutta methods analogous to those proved in the last section for spurious fixed points of these methods. Recall that a two-cycle of (3.2.1–2) is a solution sequence of the form $\mathbf{y}_{2n} = \mathbf{u}$, $\mathbf{y}_{2n+1} = \mathbf{v}$ $\forall n \geq 0$, where $\mathbf{u} \neq \mathbf{v}$.

Following Iserles *et al* [39] define the inflated method corresponding to the Runge-Kutta method (3.2.1–2) by

$$\begin{array}{c|cc} \frac{1}{2}\mathbf{c} & \frac{1}{2}\mathbf{A} & 0 \\ \hline \frac{1}{2}\mathbf{1} + \frac{1}{2}\mathbf{c} & \frac{1}{2}\mathbf{D} & \frac{1}{2}\mathbf{A} \\ \hline & \frac{1}{2}\mathbf{b}^T & \frac{1}{2}\mathbf{b}^T \end{array} \quad (4.3.1)$$

where \mathbf{A} , \mathbf{b}^T and \mathbf{c} are defined by (3.2.3)

$$D = \begin{bmatrix} \mathbf{b}^T \\ \vdots \\ \mathbf{b}^T \end{bmatrix} \quad \text{and} \quad \mathbf{1} = [1, 1, \dots, 1]^T.$$

Note that two steps of the original method with step-size h corresponds to one step of the inflated method with step-size $2h$, and this equivalence between the solution sequences of the two methods will allow us to easily extend the results of the previous section to two-cycles of Runge-Kutta methods. We begin by showing that two-cycles cannot exist for h arbitrarily small if f is globally Lipschitz.

Theorem 4.3.1 *If f is globally Lipschitz, with Lipschitz constant L , and*

$$h < \frac{1}{L(1 + \mathbf{A})(1 + \mathbf{B})} \quad (4.3.2)$$

then the Runge-Kutta method (3.2.1–2) admits no period two solutions when applied to (4.1.1).

Proof. Suppose (\mathbf{u}, \mathbf{v}) is a two-cycle of the Runge-Kutta method (3.2.1–2) with step-size h , then \mathbf{u}, \mathbf{v} are both fixed points of the inflated method (4.3.1) with step-size $2h$. Note that (4.3.2) implies that

$$2h < \frac{1}{L(1/2 + \mathbf{A}/2)(1 + \mathbf{B})}. \quad (4.3.3)$$

By Theorem 4.2.2 the inflated method does not admit any spurious fixed points if (4.3.3) holds, and hence \mathbf{u} and \mathbf{v} are both fixed points of (4.1.1). Now setting $\mathbf{y}_n = \mathbf{u}$ the existence of the two-cycle implies that there exists a solution of (3.2.1–2) such that $\mathbf{y}_{n+1} = \mathbf{v}$, and since $\mathbf{f}(\mathbf{u}) = 0$ there exists another solution of (3.2.1–2) with $\mathbf{y}_{n+1} = \mathbf{Y}_i = \mathbf{u}$ for all i , but this supplies a contradiction, since by Result 3.6.1 the solution of (4.1.1–2) is unique. ■

Note that if (\mathbf{u}, \mathbf{v}) is a two-cycle for the *explicit* Runge-Kutta method (3.2.1–2) with step-size h then \mathbf{u} and \mathbf{v} are both *spurious* fixed points of the inflated method 4.3.1 with step-size $2h$. It is clear that \mathbf{u} and \mathbf{v} are both fixed points of the inflated method. To see that they are spurious fixed points, suppose that they are not. So suppose $\mathbf{f}(\mathbf{u}) = 0$, now if $\mathbf{y}_n = \mathbf{u}$ it follows that $\mathbf{v} = \mathbf{y}_{n+1} = \mathbf{u}$, which is a fixed point, not a two-cycle. The results of the previous section can now be easily extended to two-cycles of explicit Runge-Kutta methods.

For an implicit Runge-Kutta method (3.2.1) may have more than one solution, and in this case it is not clear that \mathbf{u} and \mathbf{v} need to be spurious fixed points of the inflated method, and the results of the previous section are not so easily extended to two-cycles of implicit methods. However, it is possible to use the equivalence between the solution sequences of the Runge-Kutta method (3.2.1–2) and the corresponding inflated method (4.3.1), to extend the results of the last section. We do this by deriving results equivalent to Lemmas 4.2.5 and 4.2.6, and from which all the remaining results in this section will follow.

Lemma 4.3.2 *If $B \subset \mathbb{R}^m$ is bounded and \mathbf{f} is Lipschitz on $\mathcal{N}(B, \delta)$ with Lipschitz constant L for some $\delta > 0$, and*

$$h < \min \left(\frac{\delta}{aM}, \frac{1}{L(1 + \mathbf{A})(1 + \mathbf{B})} \right)$$

where

$$M = \sup_{\mathbf{y} \in \mathcal{N}(B, \delta)} \|\mathbf{f}(\mathbf{y})\| \quad (4.3.4)$$

then no point of B is contained in a two-cycle of (3.2.1–2), where for an implicit method we assume that the solution of (3.2.1) defined by Proposition 3.6.3 is used.

Proof. Suppose (\mathbf{u}, \mathbf{v}) is a two-cycle of the Runge-Kutta method (3.2.1–2) with step-size h and $\mathbf{u} \in B$, then \mathbf{u} is a fixed point of the inflated method (4.3.1) with step-size $2h$. By Lemma 4.2.5 $\mathbf{f}(\mathbf{u}) = 0$. Now setting $\mathbf{y}_n = \mathbf{u}$ we see that $\mathbf{y}_{n+1} = \mathbf{Y}_i = \mathbf{u}$ for all i solves the Runge-Kutta defining equations (3.2.1–2). By Proposition 3.6.3 this solution is unique, which contradicts the existence of a two-cycle. ■

Lemma 4.3.3 *If $B \subset \mathbb{R}^m$ is bounded and \mathbf{f} is continuous on $\overline{\mathcal{N}}(B, \delta)$ for some $\delta > 0$ then given $\varepsilon, \beta > 0$ there exists $H(\varepsilon, \beta) > 0$ such that if $\mathbf{u} \in B$ and (\mathbf{u}, \mathbf{v}) is a two-cycle of (3.2.1–2) with $h < H(\varepsilon, \beta)$ then*

- (i) $\max(\|\mathbf{f}(\mathbf{u})\|, \|\mathbf{f}(\mathbf{v})\|) < \varepsilon$, and,
- (ii) $\|\mathbf{u} - \mathbf{v}\| < \beta$.

Proof. Suppose (\mathbf{u}, \mathbf{v}) is a two-cycle of the Runge-Kutta method (3.2.1–2) with step-size h and $\mathbf{u} \in B$, then \mathbf{u} is a fixed point of the inflated method (4.3.1) with step-size $2h$.

By Lemma 4.2.6 there exists $H(\varepsilon) > 0$ such that if $h \in (0, H(\varepsilon))$ then $\|\mathbf{f}(\mathbf{u})\| \leq \varepsilon/2$. Let $\mathbf{y}_n = \mathbf{u}$. By Proposition 3.6.2 if $h \in (0, \delta/aM)$ where $M = \sup_{\mathbf{y} \in \mathcal{N}(B, \delta)} \|\mathbf{f}(\mathbf{y})\|$ then the Runge-Kutta defining equations (3.2.1–2) are soluble with $\mathbf{Y}_i \in \mathcal{N}(B, \delta)$ for all i .

Since $\overline{\mathcal{N}}(B, \delta)$ is compact, \mathbf{f} is uniformly continuous on $\overline{\mathcal{N}}(B, \delta)$, so given $\hat{\varepsilon} > 0$ $\exists \hat{\delta} > 0$ such that any $\mathbf{x}, \mathbf{y} \in \overline{\mathcal{N}}(B, \hat{\delta})$ with $\|\mathbf{x} - \mathbf{y}\| < \hat{\delta}$ satisfy $\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\| < \hat{\varepsilon}$. Let $\hat{\varepsilon} = \varepsilon/2$ and $\tilde{\delta} = \min(\hat{\delta}, \delta, \beta)$. Now if $h \in (0, \tilde{\delta}/BM)$ then $\|\mathbf{y}_{n+1} - \mathbf{y}_n\| < \tilde{\delta}$, since $\tilde{\delta} \leq \beta$ (ii) holds, and since $\tilde{\delta} \leq \hat{\delta}$ it follows from the uniform continuity of \mathbf{f} that $\|\mathbf{y}_{n+1} - \mathbf{y}_n\| \leq \varepsilon/2$, and hence since $\|\mathbf{f}(\mathbf{y}_n)\| \leq \varepsilon/2$ that $\|\mathbf{f}(\mathbf{y}_{n+1})\| \leq \varepsilon$, which completes the proof. ■

The results below all follow from the Lemmas 4.3.2 and 4.3.3, in a similar way to which the equivalent results were derived from Lemmas 4.2.5 and 4.2.6 in the last Section. The proofs are omitted.

Theorem 4.3.4 *If f is locally Lipschitz and period two solutions of (3.2.1–2) exist for h arbitrarily small then these tend to infinity in norm as $h \rightarrow 0$, where for an implicit method we use the solution of (3.2.1) defined by Proposition 3.2.6. By this we mean that if there exists a sequence (u_p, v_p, h_p) such that $h_p > 0 \forall p$, $h_p \rightarrow 0$ as $p \rightarrow \infty$ and (u_p, v_p) is a period two solution of the method with step-size h_p then $\|u_p\|, \|v_p\| \rightarrow \infty$ as $p \rightarrow \infty$. ■*

If we relax the condition that f is locally Lipschitz continuous and assume merely that f is continuous on \mathbb{R}^m then the following example shows that bounded spurious solutions can exist for h arbitrarily small.

Example 4.3.5 Consider the initial value problem

$$\frac{dy}{dt} = f(y), \quad \text{where } y(0) \in \mathbb{R} \quad (4.3.5)$$

where $f \in C(\mathbb{R}^m, \mathbb{R}^m)$ is defined by

$$f(y) = \begin{cases} -y^{\frac{1}{2}} & \text{if } y \geq 0 \\ (-y)^{\frac{1}{2}} & \text{if } y \leq 0. \end{cases}$$

The origin is the only fixed point of (4.3.5). Now suppose a numerical approximation is obtained using the forward Euler method (3.1.1). This yields

$$y_{n+1} = y_n + hf(y_n) \quad (4.3.6)$$

It is simple to check that $y_n = (-1)^n h^2/4$ defines a period two solution of (4.3.6) for any $h > 0$. Now suppose that the numerical solution is obtained using the inflated method (4.2.6). One step of this method with step-size h corresponds to two steps of the forward Euler method with step-size $h/2$. Thus for any $h > 0$ $y_n = h^2/16$ and $y_n = -h^2/16$ are both spurious fixed points of the method (4.2.6) for the problem (4.3.5). Notice that all the spurious solutions in this example remain bounded as $h \rightarrow 0$, and furthermore they converge to steady solutions of (4.3.5). □

The following theorem shows that if f is continuous on \mathbb{R}^m , then continuous branches of spurious period two solutions which exist for h arbitrarily small, either tend to steady solutions of the underlying differential equation, or diverge to infinity as $h \rightarrow 0$.

Theorem 4.3.6 Suppose $f \in C(\mathbb{R}^m, \mathbb{R}^m)$ and there exists a continuous branch of period two solutions $(u(h), v(h))$ of (3.2.1–2) for $h \in (0, H]$, where for an implicit method the solution of (3.2.1) is defined by Proposition 3.6.2, then as $h \rightarrow 0$ either

- (i) $\|u(h)\|$ and $\|v(h)\| \rightarrow \infty$, or,
- (ii) $\|f(u(h))\|, \|f(v(h))\|$ and $\|u(h) - v(h)\| \rightarrow 0$.

If furthermore the zeros of f are isolated then as $h \rightarrow 0$ either (i) occurs or $u(h), v(h) \rightarrow \tilde{y}$ as $h \rightarrow 0$, where \tilde{y} is a fixed point of (4.1.1). ■

The following theorem gives the same necessary condition for the bifurcation of period two solutions from \tilde{y} at $h = 0$, as was found for spurious fixed points, namely either

- (a) f is not continuous at \tilde{y} , or,
- (b) $f(\tilde{y}) = 0$ and f is not Lipschitz at \tilde{y} .

Theorem 4.3.7 Suppose there exists a sequence (u_p, v_p, h_p) such that $h_p > 0 \forall p$, $h_p \rightarrow 0$, $u_p \rightarrow \tilde{y}$ as $p \rightarrow \infty$ and (u_p, v_p) is a period two solution of (3.2.1–2) with step-size h_p for all p , and if the method is implicit then the solution of (3.2.1) is defined by Proposition 3.6.2 then if f is continuous on a neighbourhood of \tilde{y} it follows that

- (i) $v_p \rightarrow \tilde{y}$ as $p \rightarrow \infty$,
- (ii) $f(\tilde{y}) = 0$, that is \tilde{y} is a fixed point of (4.1.1), and,
- (iii) f is not Lipschitz at \tilde{y} . ■

Corollary 4.3.8 Suppose f is Lipschitz on every bounded subset of some set D , and that for an implicit method the solution of (3.2.1) defined by Proposition 3.6.3 is used, then given any positive δ, β there exists $H(\delta, \beta) > 0$ such that every point u contained in a two-cycle of (3.2.1–2) with $h < H(\delta, \beta)$, satisfies either

- i) $\inf_{x \notin D} \|u - x\| < \delta$, or,
- ii) $\|u\| > \beta$. ■

Corollary 4.3.9 Suppose f is continuous on every bounded subset of some set D , and that for an implicit method a solution of (3.2.1) defined by Proposition 3.6.2 is used, then given any positive $\varepsilon, \beta, \delta$ there exists $H(\varepsilon, \beta, \delta) > 0$ such that every point u contained in a 2-cycle of (3.2.1–2) with $h < H(\varepsilon, \beta, \delta)$ satisfies either

- i) $\inf_{x \notin D} \|u - x\| < \delta$, or,
- ii) $\|u\| > \beta$, or,
- iii) $\|f(u)\|, \|f(v)\| < \varepsilon$, and $\|u - v\| < \delta$, where v is the other point of the 2-cycle. ■

4.4 Spurious Two-Cycles of Linear Multistep Methods

The dynamics of these methods has been studied extensively. In particular:

Result 4.4.1 (Iserles [38]) *For a zero-stable linear multistep method (3.3.1), \widehat{Y} is a fixed point if and only if $f(\widehat{Y}) = 0$. ■*

Thus the method is $R^{[1]}$ and hence preserves all fixed asymptotic points of (4.1.1), and furthermore introduces no spurious steady solutions. However as with all previous methods which are $R^{[1]}$ it does not necessarily follow that the solution of the underlying system is asymptotic to the same point as the numerical solution even when the same initial value is used. No $R^{[1]}$ Runge-Kutta methods are known with order > 4 whereas we can obtain linear multistep methods of arbitrarily high order, hence these methods would seem to be very good for the long term simulation of systems (4.1.1) which are convergent to steady solutions as $T \rightarrow \infty$. The linear multistep methods which do not admit period two solutions have been studied in [39, 40, 57]. The following example shows that period two solutions can be constructed trivially if $\rho(-1) = 0$.

Example 4.4.2 (Iserles, Peplow and Stuart[39])

If $\rho(-1) = 0$ then take any f which has at least two fixed points. If $f(\tilde{y}) = f(\hat{y}) = 0$ with $\hat{y} \neq \tilde{y}$ then it is easy to check that

$$Y_n = \left(\frac{\tilde{y} + \hat{y}}{2} \right) + \left(\frac{\tilde{y} - \hat{y}}{2} \right) (-1)^n$$

is a period two solution which satisfies (3.3.1), for any $h > 0$. □

The above example prevents us from extending the results of the previous section to cover all zero-stable linear multistep methods, since the two-cycles of Example 4.4.2 exist independently of the step-size h . However if we exclude the case $\rho(-1) = 0$ we may proceed to prove similar results for period two solutions of linear multistep methods as we proved for Runge-Kutta methods. The following lemma, which shows that if the step-size is fixed then there is at most one two-cycle passing through any point of the space, provides the key to this approach.

Lemma 4.4.3 *Suppose the linear multistep method (3.3.1) is zero-stable with $\rho(-1) \neq 0$, then a two-cycle (u, v) of the method with step-size h satisfies $f(u) \neq 0$, $f(v) \neq 0$,*

$$f(u) + f(v) = 0 \tag{4.4.1}$$

and

$$\mathbf{u} = \mathbf{v} - \frac{2h\sigma(-1)\mathbf{f}(\mathbf{v})}{\rho(-1)}. \quad (4.4.2)$$

Proof. In Section 2 of [40] it is shown that a 2-cycle of a linear multistep method satisfies

$$h\sigma(1)[\mathbf{f}(\mathbf{u}) + \mathbf{f}(\mathbf{v})] = 0 \quad (4.4.3)$$

and

$$h\sigma(-1)[\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u})] = \rho(-1)[\mathbf{v} - \mathbf{u}]. \quad (4.4.4)$$

Since the method is zero-stable (4.4.1) follows from (4.4.3). Hence (4.4.4) simplifies to

$$\rho(-1)[\mathbf{v} - \mathbf{u}] = 2h\sigma(-1)\mathbf{f}(\mathbf{v}),$$

and rearranging gives (4.4.2). Finally if $\mathbf{f}(\mathbf{u}) = 0$ or $\mathbf{f}(\mathbf{v}) = 0$ then (4.4.1–2) imply that $\mathbf{u} = \mathbf{v}$, a contradiction since (\mathbf{u}, \mathbf{v}) form a two-cycle. ■

Lemma 4.4.3 allows us to explicitly classify the linear multistep methods which do not admit period two solutions.

Theorem 4.4.4

- (i) The linear multistep method (3.3.1) is not $R^{[2]}$ if $\rho(-1) = 0$.
- (ii) If $\rho(-1) \neq 0$ and the method (3.3.1) is zero-stable then it is $R^{[2]}$ if and only if $\sigma(-1) = 0$.

Remark Theorem 4.4.4 is a slight generalization of a result of Iserles *et al* [39], who proved the classification in (ii) for irreducible methods. The result of Iserles *et al* was itself a generalization of an earlier result of Stuart and Peplow [57].

Proof. (i) See Example 4.4.2.

(ii) The ‘if’ part follows from (4.4.2), since if $\sigma(-1) = 0$ then $\mathbf{u} = \mathbf{v}$ which contradicts that (\mathbf{u}, \mathbf{v}) form a 2-cycle. To prove the ‘only if’ part, take any $\mathbf{f} \in \mathcal{C}(\mathbb{R}, \mathbb{R})$ such that $\mathbf{f}(0) = -1$ and $\mathbf{f}\left(\frac{2h\sigma(-1)}{\rho(-1)}\right) = 1$. Let $\mathbf{v} = 0$ and $\mathbf{u} = \frac{2h\sigma(-1)}{\rho(-1)}$, then it is simple to check that (\mathbf{u}, \mathbf{v}) form a two-cycle. ■

Thus the class of zero-stable linear multistep methods which satisfy $\rho(-1) \neq 0$ and $\sigma(-1) = 0$ are $R^{[1,2]}$, and by considering this class of methods we can generate methods of arbitrarily high order which are $R^{[1,2]}$.

Lemma 4.4.3 also allows us to prove the following two lemmas for linear multistep methods equivalent to Lemmas 4.3.2 and 4.3.3 for Runge-Kutta methods. Note that whilst the Runge-Kutta results only hold if we use the solution of (3.2.1–2) defined by Proposition 3.6.2 or 3.6.3, the following results hold for any solution of (3.3.1). Thus in this section we do not need to make any assumptions on the solution of (3.3.1) for implicit methods.

Lemma 4.4.5 *If $B \subset \mathbb{R}^m$ is bounded and \mathbf{f} is Lipschitz on $\mathcal{N}(B, \delta)$ with Lipschitz constant L for some $\delta > 0$, then there exists $H(\delta) > 0$ such that if $h < H(\delta)$ and $\rho(-1) \neq 0$ then no point of B is contained in a 2-cycle of the zero-stable method (3.3.1).*

Proof. Let

$$M = \sup_{\mathbf{y} \in B} \|\mathbf{f}(\mathbf{y})\| \quad (4.4.5)$$

Suppose $\mathbf{v} \in B$ is contained in a two-cycle, then by Theorem 4.4.4 $\sigma(-1) \neq 0$. Hence if

$$h < \frac{\delta}{2M} \left| \frac{\rho(-1)}{\sigma(-1)} \right|$$

then (4.4.2) implies $\mathbf{u} \in \mathcal{N}(B, \delta)$. The parallelogram law states that

$$\|\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u})\|^2 + \|\mathbf{f}(\mathbf{v}) + \mathbf{f}(\mathbf{u})\|^2 = 2(\|\mathbf{f}(\mathbf{v})\|^2 + \|\mathbf{f}(\mathbf{u})\|^2). \quad (4.4.6)$$

Lipschitz continuity and (4.4.2) implies that

$$\|\mathbf{f}(\mathbf{v}) - \mathbf{f}(\mathbf{u})\| \leq 2hL \left| \frac{\sigma(-1)}{\rho(-1)} \right| \|\mathbf{f}(\mathbf{v})\|.$$

This together with (4.4.1) implies that (4.4.6) becomes

$$\|\mathbf{f}(\mathbf{v})\|^2 + \|\mathbf{f}(\mathbf{u})\|^2 \leq 2 \left(hL \frac{\sigma(-1)}{\rho(-1)} \right)^2 \|\mathbf{f}(\mathbf{v})\|^2.$$

Reversing the roles of \mathbf{u} and \mathbf{v} above, we can similarly derive that

$$\|\mathbf{f}(\mathbf{u})\|^2 + \|\mathbf{f}(\mathbf{v})\|^2 \leq 2 \left(hL \frac{\sigma(-1)}{\rho(-1)} \right)^2 \|\mathbf{f}(\mathbf{u})\|^2.$$

Now by Lemma 4.4.3 $\mathbf{f}(\mathbf{u})$ and $\mathbf{f}(\mathbf{v})$ are not both zero, so adding leads to a contradiction when

$$h < \frac{1}{L} \left| \frac{\rho(-1)}{\sigma(-1)} \right|. \quad \blacksquare$$

Lemma 4.4.6 *If $B \subset \mathbb{R}^m$ is bounded and \mathbf{f} is continuous on $\overline{\mathcal{N}}(B, \delta)$ for some $\delta > 0$, then given $\varepsilon, \beta > 0$ there exists $H(\varepsilon, \beta) > 0$ such that if $\mathbf{u} \in B$, $h < H(\varepsilon, \beta)$ and \mathbf{u} is contained in a two-cycle of the zero-stable method (3.3.1) with $\rho(-1) \neq 0$ then*

$$(i) \quad \|\mathbf{f}(\mathbf{u})\|, \|\mathbf{f}(\mathbf{v})\| < \varepsilon$$

$$(ii) \quad \|\mathbf{u} - \mathbf{v}\| < \beta$$

where \mathbf{v} is the other point of the two-cycle.

Proof. Since $\overline{\mathcal{N}}(B, \delta)$ is compact \mathbf{f} is uniformly continuous on $\overline{\mathcal{N}}(B, \delta)$. Thus given $\varepsilon > 0$ there exists $\delta_1 > 0$ such that $\forall \mathbf{x}, \mathbf{y} \in \overline{\mathcal{N}}(B, \delta)$ with $\|\mathbf{x} - \mathbf{y}\| < \delta_1$ it follows that $\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\| < \varepsilon$. Let $\delta_2 = \min(\delta, \delta_1, \beta)$.

Suppose that $\mathbf{u} \in B$ is contained in a two-cycle, then by Theorem 4.4.4 $\sigma(-1) \neq 0$. Now with M defined by (4.4.5) suppose

$$h < \frac{2\delta_2}{M} \left| \frac{\rho(-1)}{\sigma(-1)} \right|$$

then (4.4.2) implies $\|\mathbf{u} - \mathbf{v}\| < \delta_1$ and hence $\|\mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{v})\| < \varepsilon$. The result now follows from (4.4.1) and the triangle inequality. \blacksquare

The following result follows easily from Lemma 4.4.5.

Theorem 4.4.7 *If \mathbf{f} is globally Lipschitz with Lipschitz constant L , the method (3.3.1) is zero-stable with $\rho(-1) \neq 0$ and*

$$h < \frac{1}{L} \left| \frac{\rho(-1)}{\sigma(-1)} \right|$$

then the method admits no period two solutions. \blacksquare

Remark This result ties in well with the existing theory, since if $\rho(-1) = 0$ by Example 4.4.2 trivial bounded spurious solutions exist for all h and the allowed step-size

in Theorem 4.4.7 tends to zero as $\rho(-1) \rightarrow 0$, on the other hand, if $\sigma(-1) = 0$ spurious solutions cannot exist and as $\sigma(-1) \rightarrow 0$ the allowed step-size in Theorem 4.4.7 becomes unbounded.

Now Theorems 4.3.4, 4.3.6, 4.3.7 and Corollaries 4.3.8 and 4.3.9 all hold for linear multistep methods (3.3.1) which are zero-stable with $\rho(-1) \neq 0$. The proofs follow from Lemmas 4.4.5 and 4.4.6 in the same way as the Runge-Kutta results followed from Lemmas 4.3.2 and 4.3.3. Example 4.3.5 is also relevant, and shows that bounded period two solutions can exist for h arbitrarily small when f is continuous on \mathbb{R}^m .

Note that for implicit methods we have made no assumption on the scheme used to solve (3.3.1). This points out a fundamental difference between implicit linear multistep methods, for which the above results are a consequence of the method (3.3.1), and implicit Runge-Kutta methods, for which the equivalent results are a consequence of the iteration scheme used to implement the method, and which by Example 4.2.4 are not true for arbitrary solutions of the Runge-Kutta equations (3.2.1-2).

4.5 Spurious Solutions of Nonautonomous Systems

In this section we will briefly consider the solution of the nonautonomous initial value problem: find $y \in \mathbb{R}^m$ satisfying

$$\frac{dy}{dt} = f(t, y) \quad \text{for } t \geq 0 \quad \text{and} \quad y(0) = y_0 \quad (4.5.1)$$

where $f: \mathbb{R}^+ \times \mathbb{R}^m \rightarrow \mathbb{R}^m$.

If $f(t, y) = 0$ for all $t \geq 0$ then y is a fixed point of (4.5.1) and other fixed points of the numerical solution are spurious. As with autonomous systems two-cycles represent motion on the grid scale, and so must be spurious.

We begin by showing that methods which are regular for autonomous problems may admit spurious solutions for nonautonomous problems. The trapezoidal rule is $R^{[1,2]}$, but the following example shows that this method does admit spurious fixed points when applied to nonautonomous problems.

Example 4.5.1 Consider the initial value problem

$$\frac{dy}{dt} = f(t, y) = y \cos(\pi t) \quad \text{where} \quad y(0) = y_0 \in \mathbb{R}. \quad (4.5.2)$$

This has exact solution

$$y(t) = y_0 e^{\frac{1}{\pi} \sin(\pi t)}.$$

Now solve numerically using the trapezoidal rule. Note that $f(t, y) = -f(t+1, y)$, and hence if $h = 1$ then $y_n = y_0$ for all $n \geq 0$ solves the trapezoidal rule for any initial condition y_0 . \square

Note that although y_0 is a spurious fixed point of the numerical solution, the exact solution of (4.5.2) is a periodic orbit which passes through y_0 , and so we could regard the numerical solution as not being spurious, but merely as a poor approximation of the periodic orbit of the underlying system. We do not observe this behaviour for autonomous systems; in that case spurious solutions typically mark the boundary between the regions of initial conditions for which the numerical solution converges or blows up, and for autonomous systems we do not observe spurious fixed points which actually lie on periodic orbits of the underlying system.

It is clear that the spurious solution in Example 4.5.1 arises because of the oscillatory nature of f together with the special choice of h equal to half the wavelength of f , and indeed we can prove the following positive result.

Proposition 4.5.2 *Given a problem of the form (4.5.1) let*

$$p_i(\mathbf{x}) = \inf\{|s - t| : s \neq t, f_i(s, \mathbf{x}) = f_i(t, \mathbf{x}) = 0\},$$

where f_i is the i th component of \mathbf{f} and let

$$p(\mathbf{x}) = \max_i p_i(\mathbf{x}).$$

Now if $0 < h < p(\mathbf{x})$ then \mathbf{x} is not a fixed point of the trapezoidal rule for this problem.

Proof. Suppose $h < p(\mathbf{x})$ then $h < p_i(\mathbf{x})$ for some i . Now for a fixed point of the trapezoidal rule $f(t, \mathbf{x}) = -f(t+h, \mathbf{x})$. Hence $f_i(t, \mathbf{x}) = -f_i(t+h, \mathbf{x})$ and $f_i(t+nh, \mathbf{x}) = (-1)^n f_i(t, \mathbf{x})$ which provides a contradiction for some n since $h < p_i(\mathbf{x})$. \blacksquare

The following example shows that the trapezoidal rule also admits two-cycles when applied to (4.5.1). Although the solution is obviously spurious, it is more worrying than the spurious fixed point in Example 4.5.1 because in the example below \mathbf{f} is not oscillatory, and by scaling \mathbf{f} we can obtain an example of a two-cycle for h arbitrarily

small. Unlike in Example 4.5.2, there is no easily apparent natural lower bound on h below which two-cycles cannot occur.

Example 4.5.3 Consider the initial value problem

$$\frac{dy}{dt} = f(t, y) = y(t-1) \quad \text{where} \quad y(0) = y_0 \in \mathbb{R}. \quad (4.5.3)$$

This has exact solution

$$y(t) = y_0 e^{t[\frac{1}{2}-1]}.$$

Thus the origin is the only fixed point. Now solve numerically using the trapezoidal rule. Let $h = 2$ then the solution of the trapezoidal rule satisfies

$$\begin{aligned} y_{n+1} &= y_n + f(2n, y_n) + f(2(n+1), y_{n+1}) \\ y_{n+1} &= y_n + (2n-1)y_n + (2n+1)y_{n+1} \\ 2ny_{n+1} &= -2ny_n \\ y_{n+1} &= -y_n \end{aligned}$$

and hence $y_n = (-1)^n y_0$ for any initial condition y_0 . \square

It might appear at first sight that the results of the previous sections can be trivially extended to nonautonomous systems, but this is not so. Nevertheless results similar to those in the previous sections can be derived for certain nonautonomous systems. To illustrate the issues involved we will prove a result similar to Theorem 4.2.2, but for nonautonomous systems.

First note that the proof of Theorem 4.2.2 relies on Lemma 3.2.3, and this is the difficulty in extending the result to nonautonomous systems. In Lemma 3.2.3 we bound $\|f(y_n) - f(Y_i)\|$, but it is not so easy to bound $\|f(t_n, y_n) - f(t_n + c_i h, Y_i)\|$ because the function evaluations are at different times, and we will require an additional assumption on f . It would be natural to assume that f is Lipschitz in time so that

$$\|f(t_1, x) - f(t_2, x)\| \leq K|t_1 - t_2|$$

for some $K > 0$, but this is *not* sufficient to prove a satisfactory result, and we will make a stronger assumption in the lemma below.

Lemma 4.5.4 *If f is Lipschitz on $B \subseteq \mathbb{R}^m$ with Lipschitz constant L , $y_n \in B$, $Y_i \in B$ for all i , $h < 1/LA$ and*

$$\|f(t_1, x) - f(t_2, x)\| \leq K|t_1 - t_2| \cdot \inf_{t \geq 0} \|f(t, x)\| \quad (4.5.4)$$

for some $K > 0$, for all $x \in B$ and all $t_1, t_2 \geq 0$ then the solution of the Runge-Kutta defining equations (3.2.1–2) satisfies

$$\|f(t_n, y_n) - f(t_n + c_i h, Y_i)\| < \frac{(K + L)Ah}{1 - LAh} \|f(t_n, y_n)\| \quad \forall i = 1, \dots, s \quad (4.5.5)$$

where A and B are defined by (3.2.7) and (3.2.8).

Proof. Consider the solution of (3.2.1–2). Let

$$M = \max_j \|f(t_n + c_j h, Y_j)\|$$

then

$$\begin{aligned} \|f(t_n, y_n) - f(t_n + c_i h, Y_i)\| &\leq \|f(t_n, y_n) - f(t_n + c_i h, y_n)\| \\ &\quad + \|f(t_n + c_i h, y_n) - f(t_n + c_i h, Y_i)\| \\ &\leq Kh|c_i| \cdot \inf_{t \geq 0} \|f(t, y_n)\| + Lh \left\| \sum_{j=1}^s a_{ij} f(t_n + c_j h, Y_j) \right\| \\ &\leq KhA \|f(t_n, y_n)\| + LhAM. \end{aligned} \quad (4.5.6)$$

Hence

$$\|f(t_n + c_i h, Y_i)\| \leq (1 + KhA) \|f(y_n)\| + LhAM$$

for all i which implies that

$$\begin{aligned} M &\leq (1 + KhA) \|f(y_n)\| + LhAM \\ (1 - LAh)M &\leq (1 + KhA) \|f(y_n)\| \end{aligned}$$

and the result follows from (4.5.6). ■

Note that if $f(t, x) = 0$ for some $t \geq 0$ then (4.5.4) implies that $f(t, x) = 0$ for all $t \geq 0$, so that under condition (4.5.4) the fixed points of the nonautonomous system (4.5.1) are themselves fixed. Now a similar proof to that of Theorem 4.2.2 implies that

Theorem 4.5.5 *If f is globally Lipschitz, with Lipschitz constant L , (4.5.4) is satisfied for all $x \in \mathbb{R}^m$ and for all $t_1, t_2 \geq 0$, and*

$$h < \frac{1}{LA(1 + B) + KAB}, \quad (4.5.7)$$

where A and B are defined by (3.2.7) and (3.2.8), then the Runge-Kutta method (3.2.1–2) admits no spurious fixed points when applied to (4.5.1). ■

We can now go on to derive a result equivalent to Lemma 4.2.5 from which a nonautonomous version of Theorem 4.2.7 will follow, where for both results we assume that (4.5.4) holds locally; so that we allow K to depend on the set B in Lemma 4.2.5.

Finally in this chapter we note that similar results can be derived for two-cycles of Runge-Kutta and linear multistep methods. However it is not clear that nonautonomous problems which satisfy (4.5.4) are of interest, and it is not particularly instructive to state these results, and so we will not do so.

Chapter 5

Gradient Systems

5.1 Introduction

In Chapter 4 we compared and contrasted the fixed points of numerical approximations with those of the underlying system. A natural next step is to compare the dynamics of numerical approximations with the dynamics of the underlying system, for systems whose trajectories are all asymptotic to fixed points. Gradient systems have this property, and we will consider the numerical solution of these systems throughout this chapter.

Specifically we consider the numerical approximation of gradient dynamical systems defined on $U \subseteq \mathbb{R}^m$ by

$$\frac{d\mathbf{y}}{dt} = \mathbf{f}(\mathbf{y}) \quad \text{for } t \geq 0 \quad \text{and} \quad \mathbf{y}(0) = \mathbf{y}_0 \in U \quad (5.1.1)$$

where $\mathbf{y}(t) \in \mathbb{R}^m$ and $\mathbf{f}: U \rightarrow \mathbb{R}^m$ is locally Lipschitz and satisfies

$$\mathbf{f}(\mathbf{y}) = -\nabla F(\mathbf{y}) \quad (5.1.2)$$

where $F \in \mathcal{C}^1(U, \mathbb{R})$ is bounded below and satisfies

$$F(\mathbf{y}) \rightarrow \infty \quad \text{as} \quad \|\mathbf{y}\| \rightarrow \infty. \quad (5.1.3)$$

Recall from Section 2.4 that such a system is a gradient system and that (assuming the fixed points of \mathbf{f} are isolated) this implies that the solution $\mathbf{y}(t)$ of (5.1.1–2) is asymptotic to a fixed point of the system for any initial condition \mathbf{y}_0 .

It might at first seem that gradient systems should be of little interest since the possible dynamics of these systems is so simple, but this is not so for a variety of reasons.

Scalar reaction-diffusion equations are partial differential equations which are in gradient form [17]. The Cahn-Hilliard equation, which models the process of coarsening in solid phase separation, is also an important example of a partial differential equation which is in gradient form [16]. Under suitable spatial discretization these systems generate gradient systems of a similar form to (5.1.1–2) [16, 17], and the study of these systems is worthwhile for this reason alone. Furthermore gradient systems have also been fundamental in the development of many concepts in the theory of ordinary differential equations, and are also important for this reason.

Finally from a numerical analysis of dynamical systems viewpoint gradient systems are of interest because they are *not* chaotic. Our ultimate objective is a theory for the numerical solution of chaotic systems, however when numerical solutions display chaotic behaviour it is important to determine if the underlying system is chaotic, or whether the chaos is numerically generated. Since we know that every trajectory of (5.1.1–2) is asymptotic to a fixed point, if a numerical approximation of such a system displayed apparently chaotic behaviour or some other complex dynamical feature then this would cast grave doubt on any chaotic computations produced using that particular method.

In Section 5.2 we consider (5.1.1–2) under the assumption that \mathbf{f} is either locally or globally Lipschitz. In the case where \mathbf{f} is globally Lipschitz, in Theorem 5.2.3, we will show that the numerical solution generated by any Runge-Kutta method defines a continuous discrete gradient system with the same fixed points and Lyapunov functional as (5.1.1–2) if the step-size h is sufficiently small, and we derive a sufficient bound on the step-size. The key to proving this result is Proposition 5.2.2 in which we show that when \mathbf{f} is Lipschitz the numerical solution of (5.1.1–2) by a Runge-Kutta method satisfies an inequality of the form

$$F(\mathbf{y}_{n+1}) - F(\mathbf{y}_n) \leq \gamma(h) \|\mathbf{y}_{n+1} - \mathbf{y}_n\|^2$$

where $\gamma(\bullet)$ is independent of n , and is negative for h sufficiently small.

When \mathbf{f} is locally Lipschitz we cannot preserve the gradient structure globally for arbitrary Runge-Kutta methods. However, we note that the set $B = \{\mathbf{x}: F(\mathbf{x}) \leq w\}$ is forward invariant under the evolution of (5.1.1–2) for any w , and then, using Proposi-

tion 5.2.2 again, prove that B is also forward invariant for the numerical approximation if h is sufficiently small. Indeed, in Theorem 5.2.5, we show that for h sufficiently small the numerical solution defines a continuous discrete gradient system on B with the same fixed points and Lyapunov functional as (5.1.1–2) on B .

In Section 5.3 we consider (5.1.1–2) under the assumption that \mathbf{f} satisfies a one-sided Lipschitz condition on \mathbb{R}^m

$$\langle \mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle \leq c \|\mathbf{u} - \mathbf{v}\|^2 \quad \forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^m, \quad (5.1.4)$$

with $c > 0$, and prove that for $h < 1/c$ and $\theta \in [1/2, 1]$ both the one and two-stage theta methods define continuous discrete gradient systems on \mathbb{R}^m with the same fixed points as (5.1.1–2). However, unlike in Section 5.2, when \mathbf{f} merely satisfies a one-sided Lipschitz condition the Lyapunov functional is not the same as that of (5.1.1–2) but is a perturbation of it. Related results can be found in Elliott and Stuart [17] where it is shown that solution of a class of gradient systems which satisfy a one-sided Lipschitz condition by any of the first three backward differentiation formulae defines a continuous discrete gradient system with the same fixed points as the underlying system and with a Lyapunov functional which is a perturbation of the Lyapunov functional of the underlying system.

5.2 Approximation of Lipschitz Gradient Systems

In this section we consider the numerical solution of the gradient system (5.1.1–2) under the assumption that \mathbf{f} is globally or locally Lipschitz. We will show that every Runge-Kutta method preserves the underlying gradient structure on \mathbb{R}^m for h sufficiently small if \mathbf{f} is globally Lipschitz, whilst if \mathbf{f} is locally Lipschitz then the gradient structure is preserved on any bounded set for h sufficiently small.

First we consider the solution of (5.1.1–2) by the forward Euler method when \mathbf{f} is globally Lipschitz.

Theorem 5.2.1 *If \mathbf{f} is globally Lipschitz on \mathbb{R}^m with Lipschitz constant L and the gradient system (5.1.1–2) is approximated numerically by the forward Euler method (3.1.1) with $h < 1/L$ then the numerical solution defines a continuous discrete gradient system on \mathbb{R}^m which has the same fixed points and the same Lyapunov functional as (5.1.1–2).*

Proof. Since the method is explicit the numerical solution trivially defines a discrete dynamical system and continuity with respect to initial data follows as in the proof of Theorem 3.6.4. Now Theorem 4.2.1 implies that (5.1.1-2) and the discrete dynamical system defined by its numerical approximation have the same fixed points.

To establish that this system is in gradient form let $F_h(\mathbf{y}) = F(\mathbf{y})$ and note that (i) and (ii) of Definition 2.5.3 thus hold automatically. Now (2.4.4) implies that

$$\begin{aligned} F_h(\mathbf{y}_{n+1}) - F_h(\mathbf{y}_n) &\leq \langle \mathbf{f}(\mathbf{y}_n), \mathbf{y}_n - \mathbf{y}_{n+1} \rangle + L\|\mathbf{y}_{n+1} - \mathbf{y}_n\|^2 \\ &= \left(L - \frac{1}{h}\right) \|\mathbf{y}_{n+1} - \mathbf{y}_n\|^2 \end{aligned}$$

since $\mathbf{f}(\mathbf{y}_n) = \frac{1}{h}(\mathbf{y}_{n+1} - \mathbf{y}_n)$. Thus for $h < 1/L$ either $F_h(\mathbf{y}_{n+1}) < F_h(\mathbf{y}_n)$ or $\mathbf{y}_{n+1} = \mathbf{y}_n$, and (iii) and (iv) of Definition 2.5.3 follow. Thus the discrete system is in gradient form and has the same Lyapunov functional as (5.1.1-2). ■

Recall that if $\mathbf{A} = 0$ then (3.2.15) holds and the Runge-Kutta method gives a solution sequence equivalent to that of the forward Euler method, and hence by Theorem 5.2.1 defines a continuous discrete gradient system with the same fixed points and Lyapunov functional as (5.1.1-2) if \mathbf{f} is globally Lipschitz. We can prove a similar result for other Runge-Kutta methods, but first we need a preparatory proposition.

Proposition 5.2.2 *If \mathbf{f} is defined by (5.1.2) is Lipschitz on a convex set $B \subseteq \mathbb{R}^m$ with Lipschitz constant L , $\mathbf{y}_n \in B$ and either $\mathbf{A} = 0$ or*

$$h < \frac{1}{L\mathbf{A}(1 + \mathbf{B})}, \quad (5.2.1)$$

then any solution of the Runge-Kutta defining equations (3.2.1-2) which satisfies $\mathbf{Y}_i \in B$ for all i and $\mathbf{y}_{n+1} \in B$ also satisfies

$$F(\mathbf{y}_{n+1}) - F(\mathbf{y}_n) \leq \left[L - \frac{1}{h} + \frac{L\mathbf{A}\mathbf{B}}{1 - Lh\mathbf{A}(1 + \mathbf{B})} \right] \|\mathbf{y}_{n+1} - \mathbf{y}_n\|^2 \quad (5.2.2)$$

where \mathbf{A} and \mathbf{B} are defined by (3.2.7) and (3.2.8).

Proof. Equation (2.4.4) implies that

$$\begin{aligned} F(\mathbf{y}_{n+1}) - F(\mathbf{y}_n) &\leq \langle \mathbf{f}(\mathbf{y}_n), \mathbf{y}_n - \mathbf{y}_{n+1} \rangle + L\|\mathbf{y}_{n+1} - \mathbf{y}_n\|^2 \\ &= \left(L - \frac{1}{h}\right) \|\mathbf{y}_{n+1} - \mathbf{y}_n\|^2 + \langle \mathbf{f}(\mathbf{y}_n) + \frac{1}{h}(\mathbf{y}_n - \mathbf{y}_{n+1}), \mathbf{y}_n - \mathbf{y}_{n+1} \rangle. \end{aligned}$$

Notice that

$$\begin{aligned} \frac{1}{h}(\mathbf{y}_{n+1} - \mathbf{y}_n) - \mathbf{f}(\mathbf{y}_n) &= \sum_{i=1}^s b_i \mathbf{f}(\mathbf{Y}_i) - \mathbf{f}(\mathbf{y}_n) \\ &= \sum_{i=1}^s b_i (\mathbf{f}(\mathbf{Y}_i) - \mathbf{f}(\mathbf{y}_n)) \end{aligned}$$

and hence

$$F(\mathbf{y}_{n+1}) - F(\mathbf{y}_n) \leq \left(L - \frac{1}{h}\right) \|\mathbf{y}_{n+1} - \mathbf{y}_n\|^2 + \|\mathbf{y}_{n+1} - \mathbf{y}_n\| \cdot \|\mathbf{f}(\mathbf{y}_n) - \sum_{i=1}^s b_i \mathbf{f}(\mathbf{Y}_i)\|$$

and the result follows trivially from (3.2.15) if $\mathbf{A} = 0$, and from (3.2.19) if $\mathbf{A} \neq 0$. ■

Define $\gamma: \left(0, \frac{1}{L\mathbf{A}(1+\mathbf{B})}\right) \rightarrow \mathbb{R}$ by

$$\gamma(h) = L - \frac{1}{h} + \frac{L\mathbf{A}\mathbf{B}}{1 - Lh\mathbf{A}(1+\mathbf{B})}. \quad (5.2.3)$$

Note that $\gamma(h)$ is continuous and strictly increasing for $h \in \left(0, \frac{1}{L\mathbf{A}(1+\mathbf{B})}\right)$, $\gamma(h) \rightarrow -\infty$ as $h \rightarrow 0$ and $\gamma(h) \rightarrow +\infty$ as $h \rightarrow \frac{1}{L\mathbf{A}(1+\mathbf{B})}$. Hence γ has a unique zero $h_0 \in \left(0, \frac{1}{L\mathbf{A}(1+\mathbf{B})}\right)$ and $\gamma(h) < 0$ for $h \in (0, h_0)$. Now it follows from (5.2.2) that if $h \in (0, h_0)$ and the conditions of Proposition 5.2.2 are satisfied then either $F(\mathbf{y}_{n+1}) < F(\mathbf{y}_n)$ or $\mathbf{y}_{n+1} = \mathbf{y}_n$. This is the basis of the proof of the following theorem.

Theorem 5.2.3 *If \mathbf{f} is globally Lipschitz on \mathbb{R}^m with Lipschitz constant L and the gradient system (5.1.1–2) is approximated numerically by a Runge-Kutta method (3.2.1–2) with $\mathbf{A} > 0$ and*

$$h < h_0 < \frac{1}{L\mathbf{A}(1+\mathbf{B})} \quad (5.2.4)$$

where h_0 is the zero of γ described above then the numerical solution defines a continuous discrete gradient system on \mathbb{R}^m which has the same fixed points and the same Lyapunov functional as (5.1.1–2).

Proof. Note that (5.2.4) and (3.2.10) imply that $h < 1/La$ and hence Result 3.6.1 implies that the numerical solution defines a discrete dynamical system on \mathbb{R}^m . Continuity with respect to initial data follows as in the proof of Theorem 3.6.4. Now Theorem 4.2.2 implies that (5.1.1–2) and the discrete dynamical system defined by its numerical approximation have the same fixed points.

To establish that this system is in gradient form let $F_h(\mathbf{y}) = F(\mathbf{y})$ and note that (i) and (ii) of Definition 2.5.3 hold automatically. Since $h < h_0$ we can apply Proposition 5.2.2 and (5.2.2) implies that either $F_h(\mathbf{y}_{n+1}) < F(\mathbf{y}_n)$ or $\mathbf{y}_{n+1} = \mathbf{y}_n$ and (iii) and (iv) of Definition 2.5.3 also follow. Thus the discrete system is in gradient form and has the same Lyapunov functional as (5.1.1–2). ■

The proof of Theorem 5.2.3 uses Proposition 5.2.2 which is in turn proved using inequality (3.2.19). Since (3.2.15) holds for the forward Euler method we are effectively treating all other Runge-Kutta methods as perturbations of the forward Euler method in the proof of Theorem 5.2.3, and hence the step-size bound given by this theorem may not be optimal for other methods. Indeed we shall now demonstrate this for the two-stage theta method (3.5.2). Note that $\mathbf{A} = \mathbf{B} = \mathbf{1}$ for this method and this implies that $h_0 = \frac{1}{L} \left(1 - \frac{1}{\sqrt{2}}\right)$ in (5.2.4). The following result improves this step-size bound to $h_0 = 1/L$.

Proposition 5.2.4 *If \mathbf{f} is globally Lipschitz with Lipschitz constant L and the gradient system (5.1.1–2) is approximated numerically using the two-stage theta method (3.5.2) with $h < 1/L$ then the numerical solution defines a continuous discrete gradient system which has the same Lyapunov functional and the same fixed points as (5.1.1–2).*

Proof. By Result 3.6.1 the numerical solution defines a discrete dynamical system for $h < 1/L$, and continuity with respect to initial data follows as in the proof of Theorem 3.6.4. Theorem 4.4.1 implies that (5.1.1–2) and the discrete system defined by the numerical approximation have the same fixed points.

It remains to show that the gradient structure is preserved. Let $F_h(\mathbf{y}) = F(\mathbf{y})$, then (i) and (ii) of Definition 2.5.3 hold automatically. By (2.4.2)

$$F_h(\mathbf{y}_{n+1}) - F_h(\mathbf{y}_n) \leq \langle \mathbf{f}(\mathbf{y}_{n+1}), \mathbf{y}_n - \mathbf{y}_{n+1} \rangle + L\|\mathbf{y}_{n+1} - \mathbf{y}_n\|^2 \quad (5.2.5)$$

and by (2.4.4)

$$F_h(\mathbf{y}_{n+1}) - F_h(\mathbf{y}_n) \leq \langle \mathbf{f}(\mathbf{y}_n), \mathbf{y}_n - \mathbf{y}_{n+1} \rangle + L\|\mathbf{y}_{n+1} - \mathbf{y}_n\|^2. \quad (5.2.6)$$

Now adding $(1 - \theta)(5.2.5) + \theta(5.2.6)$ implies

$$F_h(\mathbf{y}_{n+1}) - F_h(\mathbf{y}_n) \leq \langle \theta \mathbf{f}(\mathbf{y}_n) + (1 - \theta) \mathbf{f}(\mathbf{y}_{n+1}), \mathbf{y}_n - \mathbf{y}_{n+1} \rangle + L\|\mathbf{y}_{n+1} - \mathbf{y}_n\|^2$$

$$\begin{aligned}
&= \frac{1}{h} \langle \mathbf{y}_{n+1} - \mathbf{y}_n, \mathbf{y}_n - \mathbf{y}_{n+1} \rangle + L \|\mathbf{y}_{n+1} - \mathbf{y}_n\|^2 \\
&= \left(L - \frac{1}{h} \right) \|\mathbf{y}_{n+1} - \mathbf{y}_n\|^2.
\end{aligned}$$

Thus for $h < 1/L$ we have $F_h(\mathbf{y}_{n+1}) \leq F_h(\mathbf{y}_n)$ and $F_h(\mathbf{y}_{n+1}) = F_h(\mathbf{y}_n)$ if and only if $\mathbf{y}_{n+1} = \mathbf{y}_n$. Hence (iii) and (iv) of Definition 2.5.3 also hold, and the discrete dynamical system is in gradient form, with the same Lyapunov functional as the underlying system. ■

We now relax the condition that \mathbf{f} is globally Lipschitz and consider (5.1.1–2) under the assumption that \mathbf{f} is locally Lipschitz. Let

$$B = \{\mathbf{x}: F(\mathbf{x}) \leq w\} \quad (5.2.7)$$

for some $w > \inf_{\mathbf{x} \in \mathbb{R}^m} F(\mathbf{x})$. Note that B is bounded by (5.1.3). Since $F(\mathbf{y}(t))$ is non-increasing for any solution trajectory $\mathbf{y}(t)$ of (5.1.1–2) it follows that B is forward invariant under the evolution of (5.1.1–2), and hence that the restriction of (5.1.1–2) to B defines a gradient system on B . The final theorem in this section shows that the numerical solution also preserves the gradient structure and defines a discrete gradient system on B for h sufficiently small.

Theorem 5.2.5 *Suppose that \mathbf{f} is locally Lipschitz and (5.1.1–2) is approximated numerically by a Runge-Kutta method, where for an implicit method the solution of (3.2.1–2) is defined by Proposition 3.6.3. Let*

$$B = \{\mathbf{x}: F(\mathbf{x}) \leq w\}$$

for any $w > \inf_{\mathbf{x} \in \mathbb{R}^m} F(\mathbf{x})$ then there exists $h_w > 0$ such that if $h \in (0, h_w)$ then the numerical solution defines a continuous discrete dynamical system on B which has the same Lyapunov functional and fixed points as the restriction of the gradient system (5.1.1–2) to B .

Proof. Pick arbitrary $\varepsilon > 0$ and suppose that h satisfies (3.6.7) then for $\mathbf{y}_n \in B$ Proposition 3.6.3 defines a unique solution of (3.2.1–2) which satisfies $\mathbf{Y}_i \in B(\mathbf{y}_n, \varepsilon)$ for all i . If in addition we assume that $h < \varepsilon/BM$ where M is defined by (3.6.3) then

$$\|\mathbf{y}_{n+1} - \mathbf{y}_n\| = h \left\| \sum_{i=1}^s b_i \mathbf{f}(\mathbf{Y}_i) \right\|$$

$$\leq hBM$$

$$< \varepsilon.$$

Hence $\mathbf{y}_{n+1} \in B(\mathbf{y}_n, \varepsilon)$ and $\mathbf{Y}_i \in B(\mathbf{y}_n, \varepsilon)$ for all i and so Proposition 5.2.2 applies and (5.2.2) holds. Now assuming $h < h_0$ where $h_0 \in (0, \frac{1}{L\mathbf{M}(1+B)})$ is the zero of $\gamma(\bullet)$, as defined by (5.2.3), then (5.2.2) implies that either $F(\mathbf{y}_{n+1}) < F(\mathbf{y}_n)$ or $\mathbf{y}_{n+1} = \mathbf{y}_n$, and in either case $F(\mathbf{y}_{n+1}) \leq F(\mathbf{y}_n)$ and so $\mathbf{y}_{n+1} \in B$. Now Theorem 3.6.4 implies that the numerical solution defines a continuous discrete dynamical system on B . Since either $F(\mathbf{y}_{n+1}) < F(\mathbf{y}_n)$ or $\mathbf{y}_{n+1} = \mathbf{y}_n$ it follows that the system is in gradient form with Lyapunov functional $F_h \equiv F|_B$. Finally Lemma 4.2.5 implies that this discrete gradient system has the same fixed points as the restriction of the gradient system (5.1.1–2) to B . ■

Remark It is important to note that the numerical solution not only preserves the Lyapunov functional and the fixed points of the underlying system on B , but that B is forward invariant for the numerical solution, so that the gradient structure of the numerical solution on B cannot be destroyed by having trajectories which escape from B .

5.3 One-sided Lipschitz Gradient Systems

In this section we consider the numerical solution of (5.1.1–2) under the assumption that \mathbf{f} satisfies a one-sided Lipschitz condition (5.1.4). We would like to preserve the gradient structure globally on \mathbb{R}^m without having to assume that \mathbf{f} is globally Lipschitz as we did in Theorem 5.2.3. We can do this for both the one- and two-stage theta methods if $\theta \in [1/2, 1]$ and \mathbf{f} satisfies a one-sided Lipschitz condition (5.1.4). We prove the result first for the two-stage theta method.

Theorem 5.3.1 *If \mathbf{f} a one-sided Lipschitz condition (5.1.4) and the gradient system (5.1.1–2) is approximated numerically using the two-stage theta method (3.5.2) with $\theta \in [1/2, 1]$ and $h \in (0, 1/c)$ then the numerical solution defines a continuous discrete gradient system on \mathbb{R}^m with Lyapunov functional $F_h(\bullet)$ given by*

$$F_h(\mathbf{y}) = F(\mathbf{y}) + \frac{h}{2}(1 - \theta)\|\mathbf{f}(\mathbf{y})\|^2 \quad (5.3.1)$$

and the same fixed points as (5.1.1–2).

Proof. Result 3.6.16 implies that the equations defining the two-stage theta method (3.5.2) are uniquely soluble for $h \in (0, 1/c)$ and hence that the numerical solution defines a discrete dynamical system. Continuity with respect to initial data follows from Theorem 14.3 of [27].

Theorem 4.4.1 implies that the fixed points of this discrete dynamical system are the same as those of (5.1.1–2). To show that the discrete system is in gradient form consider (5.3.1). Note that $F_h(\mathbf{y}) \geq F(\mathbf{y})$ and so (i) and (ii) of Definition 2.5.3 hold automatically. Now (2.4.2) implies that

$$\begin{aligned}
 F(\mathbf{y}_{n+1}) - F(\mathbf{y}_n) &\leq \langle \mathbf{f}(\mathbf{y}_{n+1}), \mathbf{y}_n - \mathbf{y}_{n+1} \rangle + c\|\mathbf{y}_{n+1} - \mathbf{y}_n\|^2 \\
 &= \left\langle \frac{1}{h} [\mathbf{y}_{n+1} - \mathbf{y}_n - h(1-\theta)\mathbf{f}(\mathbf{y}_n) - h\theta\mathbf{f}(\mathbf{y}_{n+1})] + \mathbf{f}(\mathbf{y}_{n+1}), \mathbf{y}_n - \mathbf{y}_{n+1} \right\rangle \\
 &\quad + c\|\mathbf{y}_{n+1} - \mathbf{y}_n\|^2 \\
 &= \left(c - \frac{1}{h} \right) \|\mathbf{y}_{n+1} - \mathbf{y}_n\|^2 + (1-\theta) \langle \mathbf{f}(\mathbf{y}_n) - \mathbf{f}(\mathbf{y}_{n+1}), \mathbf{y}_{n+1} - \mathbf{y}_n \rangle \\
 &= \left(c - \frac{1}{h} \right) \|\mathbf{y}_{n+1} - \mathbf{y}_n\|^2 \\
 &\quad + h(1-\theta) \langle \mathbf{f}(\mathbf{y}_n) - \mathbf{f}(\mathbf{y}_{n+1}), (1-\theta)\mathbf{f}(\mathbf{y}_n) + \theta\mathbf{f}(\mathbf{y}_{n+1}) \rangle \\
 &= \left(c - \frac{1}{h} \right) \|\mathbf{y}_{n+1} - \mathbf{y}_n\|^2 + \frac{h}{2}(1-\theta) [\|\mathbf{f}(\mathbf{y}_n)\|^2 - \|\mathbf{f}(\mathbf{y}_{n+1})\|^2] \\
 &\quad - \frac{h}{2}(1-\theta)(2\theta-1) \|\mathbf{f}(\mathbf{y}_n) - \mathbf{f}(\mathbf{y}_{n+1})\|^2.
 \end{aligned}$$

Thus

$$F_h(\mathbf{y}_{n+1}) - F_h(\mathbf{y}_n) \leq \left(c - \frac{1}{h} \right) \|\mathbf{y}_{n+1} - \mathbf{y}_n\|^2 - \frac{h}{2}(1-\theta)(2\theta-1) \|\mathbf{f}(\mathbf{y}_n) - \mathbf{f}(\mathbf{y}_{n+1})\|^2$$

and so for $\theta \in [1/2, 1]$

$$F_h(\mathbf{y}_{n+1}) - F_h(\mathbf{y}_n) \leq \left(c - \frac{1}{h} \right) \|\mathbf{y}_{n+1} - \mathbf{y}_n\|^2.$$

Thus $F_h(\mathbf{y}_{n+1}) < F_h(\mathbf{y}_n)$ unless \mathbf{y}_n is a fixed point of the discrete dynamical system and so (iii) and (iv) of Definition 2.5.3 hold, and the system is in gradient form. ■

Remark Unlike the results in the previous section the Lyapunov functional F_h (5.3.1) of the discrete gradient system defined by the two-stage theta method is not the same as the Lyapunov functional F of the underlying system (5.1.1–2), but is a perturbation of

it. Note that not only are the fixed points of (5.1.1–2) and its discrete counterpart the same, but that F and F_h agree at these points. Similar results are obtained in Elliott and Stuart [17] for the first three backward differentiation formulae; these methods are shown to preserve the gradient structure globally under a perturbation of the Lyapunov functional when f satisfies a one-sided Lipschitz condition.

We can prove a similar result for the one-stage theta method (3.5.1) using the relationship between this method and the two-stage theta method. We write (3.5.1) as

$$Y^n = y_n + h\theta f(Y^n) \quad (5.3.2)$$

$$y_{n+1} = y_n + hf(Y^n) \quad (5.3.3)$$

to show the dependence of the stage value Y on n . With this notation we present the final result of this chapter.

Corollary 5.3.2 *If f satisfies a one-sided Lipschitz condition (5.1.4) and the gradient system (5.1.1–2) is approximated numerically using the one-stage theta method (5.3.2–3) with $\theta \in [1/2, 1]$ and $h \in (0, 1/c)$ then the numerical solution defines a continuous discrete gradient system on \mathbb{R}^m with Lyapunov functional $F_h(\bullet)$ given by*

$$F_h(y_n) = F(Y^n) + \frac{h}{2}(1 - \theta)\|f(Y^n)\|^2 \quad (5.3.4)$$

and the same fixed points as (5.1.1–2).

Proof. Result 3.6.15 implies that the equations defining the one-stage theta method are uniquely soluble for $h \in (0, 1/c)$ and hence that the numerical solution defines a discrete dynamical system. Continuity with respect to initial data follows from Theorem 14.3 of [27].

Lemma 2 of Hairer *et al* [25] implies that the one-stage theta method is $R^{[1]}$, and hence that the fixed points of the discrete dynamical system defined by the numerical approximation are the same as those of (5.1.1–2).

Note that since (5.3.3) is uniquely soluble for any $y_n \in \mathbb{R}^m$ and any $h \in (0, 1/c)$, $F_h(y)$ is well defined by (5.3.4). F_h clearly satisfies (i) of Definition 2.5.3.

To show that F_h satisfies (ii) of Definition 2.5.3 from (5.3.4) it is sufficient to show that $\|Y\| \rightarrow \infty$ as $\|y\| \rightarrow \infty$ where

$$Y = y + h\theta f(Y). \quad (5.3.5)$$

But (5.3.5) implies that

$$\|Y - y\|^2 = h\theta \|f(Y)\|^2. \quad (5.3.6)$$

Now suppose that $\|y\| \rightarrow \infty$ but that $\|Y\|$ remains bounded. Then the left-hand side of (5.3.6) becomes unbounded, but by continuity of f the right-hand side remains bounded, which supplies the required contradiction. Hence F_h satisfies (ii) of Definition 2.5.3.

To show that F_h satisfies (iii) and (iv) of Definition 2.5.3 we will exploit the close relationship between the one- and two-stage theta methods which was established in Example 3.5.1. Note that

$$Y^{n+1} = y_{n+1} + h\theta f(Y^{n+1})$$

and hence

$$\begin{aligned} Y^{n+1} - Y^n &= y_{n+1} - y_n + h\theta [f(Y^{n+1}) - f(Y^n)] \\ &= h(1 - \theta)f(Y^n) + h\theta f(Y^{n+1}). \end{aligned} \quad (5.3.7)$$

Thus the stage values of the one-stage theta method at successive steps satisfy the two-stage theta method, and Theorem 5.3.1 implies that $F_h(y_{n+1}) \leq F_h(y_n)$, and $F_h(y_{n+1}) = F_h(y_n)$ if and only if $Y^{n+1} = Y^n$. Now if $Y^{n+1} = Y^n$ then by (5.3.7) $f(Y^n) = 0$ and (5.3.3) implies that $y_{n+1} = y_n$. Thus $F_h(y_{n+1}) < F_h(y_n)$ unless $y_{n+1} = y_n$ and (iii) and (iv) of Definition 2.5.3 follow and the discrete system is in gradient form. ■

Remark In both Theorem 5.3.1 and Corollary 5.3.2 to preserve the gradient structure we require $h < 1/c$ where c is the one-sided Lipschitz constant. In Theorem 2.4.4 in the previous section we required $h < h_0$ to preserve the gradient structure and it follows from (5.2.3) that $h_0 \leq 1/L$. Thus we require $h < 1/c$ or $h < 1/L$ for the relevant Lipschitz constant in every result in this chapter, and this seems to be a necessary

bound on the step-size in order to preserve the gradient structure of the underlying system.

Chapter 6

Dissipative Systems

6.1 Introduction

In the last chapter we considered dynamical systems for which all trajectories are asymptotic to fixed points, and compared the dynamics of the numerical approximation with the dynamics of the underlying system. In this chapter we will generalize our theory by considering dynamical systems for which trajectories need not be asymptotic to fixed points, but which possess a bounded *absorbing set* which all trajectories enter in a finite time and thereafter remain inside. Recall from Definition 2.2.1 that such systems are said to be *dissipative*. Clearly the asymptotic behaviour of the system must be confined to the absorbing set, but it is worth emphasising that the dynamics within this set may be very complicated, and indeed many chaotic nonlinear systems are dissipative. We will seek to establish conditions under which the numerical solution is also dissipative, since if the absorbing set is destroyed by the discretization then incorrect asymptotic behaviour will be observed for at least some initial conditions.

We consider the numerical approximation of dynamical systems defined by

$$\frac{d\mathbf{y}}{dt} = \mathbf{f}(\mathbf{y}) \tag{6.1.1}$$

for $t \in [0, \infty)$, $\mathbf{y}(t) \in \mathbb{R}^m$ and arbitrary initial condition $\mathbf{y}(0) = \mathbf{y}_0$. We will assume that $\mathbf{f}: \mathbb{R}^m \rightarrow \mathbb{R}^m$ is locally Lipschitz, and for most of this chapter we will make the additional structural assumption on \mathbf{f} that

$$\langle \mathbf{f}(\mathbf{y}), \mathbf{y} \rangle \leq \alpha - \beta \|\mathbf{y}\|^2 \tag{6.1.2}$$

for some $\alpha \geq 0$ and $\beta > 0$. Recall from Section 2.2.1 that (6.1.1–2) does define a dissipative dynamical system and that the ball $B(0, R)$ is an absorbing set for any radius $R > \sqrt{\alpha/\beta}$.

Systems of the form (6.1.1–2) arise in many applications, but often through spatial discretization of partial differential equations. The Cahn-Hilliard equation, the Navier-Stokes equations in two dimensions, the complex Ginzburg-Landau equation, and the Kuramoto-Sivashinsky equation all satisfy an infinite dimensional analogue of (6.1.2) [58]. Under suitable spatial discretization they generate systems of the form (6.1.1–2). For example Foias and Titi [19] derive a finite difference approximation to the Kuramoto-Sivashinsky equation, Elliott and Stuart [17] derive finite difference and finite element approximations to a class of semi-linear reaction-diffusion equations, and Lord [43] discretizes the complex Ginzburg-Landau equation; all of these spatial discretizations define dissipative dynamical systems of the form (6.1.1–2).

Although systems of this form are all dissipative and their asymptotic behaviour is confined to a bounded absorbing set, we emphasise that these systems can display a variety of interesting dynamical features ranging from multiple competing equilibria (the Cahn-Hilliard equation) through periodic and quasi-periodic behaviour (the complex Ginzburg-Landau equation) to chaos (the Kuramoto-Sivashinsky equation).

The Lorenz equations (2.2.5) represent another important example of a dissipative dynamical system which displays apparently chaotic dynamics. Recall from Example 2.2.3 that this system is of the form (6.1.1–2) after translation of the coordinate system.

In Section 6.2 we consider the dynamics of numerical solutions to (6.1.1–2) generated by algebraically stable Runge-Kutta methods, and would like to show that the numerical solution defines a dissipative discrete dynamical system. Theorem 3.6.18 shows that for a DJ-irreducible algebraically stable method with invertible A applied to (6.1.1–2) the Runge-Kutta defining equations (3.2.1–2) are always soluble and hence that there exists a solution sequence $\{\mathbf{y}_n\}_{n=0}^{\infty}$ for any initial condition \mathbf{y}_0 and any step-size $h > 0$. However, recall from Example 3.6.19 that (3.2.1–2) may admit multiple solutions and hence that the solution sequence $\{\mathbf{y}_n\}_{n=0}^{\infty}$ need not be unique. This implies that in general the numerical solution does not even define a discrete dynamical system, and so we cannot consider whether it is dissipative or not. This leads us to consider the dissipativity of the numerical solution in terms of the generalized concept

of dissipativity for multi-valued maps from Section 2.5.2.

To prove the existence of an absorbing set often requires a step-size bound which is dependent on the initial data; however absorbing sets with step-size bounds independent of the initial data have been constructed in [17, 18, 19, 43] for spatial semi-discretizations of partial differential equations satisfying infinite dimensional analogues of (6.1.2) *and* for temporal discretization of the resulting ordinary differential equations (which are of the form (6.1.1–2)). We present an example which shows that when (6.1.1–2) is approximated by a non A-stable method then a step-size bound dependent on initial data will be required. This leads us to consider the numerical solution of (6.1.1–2) by algebraically stable Runge-Kutta methods, and the major result in this chapter is Theorem 6.2.2 in which we show that the numerical approximation defined by any algebraically stable Runge-Kutta method is dissipative for *any* fixed step-size $h > 0$. Recall that there exist algebraically stable Runge-Kutta methods of arbitrarily high order, and hence that we can approximate (6.1.1–2) using methods of arbitrarily high order, whilst still retaining the dissipativity of the underlying system. The full discretizations of the Kuramoto-Sivashinsky equation and reaction diffusion equations considered in [18] and [17] respectively are all of first or second order in time and often correspond to applying the backward Euler method to the appropriate semi-discretized system, although explicit and mixed explicit/implicit treatment of the nonlinear part of \mathbf{f} is also considered in both papers. Theorem 6.2.2 shows that not only the backward Euler method but any algebraically stable Runge-Kutta method can be used to solve these semi-discretized systems whilst retaining the dissipativity of the underlying system and that special treatment of the nonlinear part of \mathbf{f} is not necessary (although it may still be desirable for computational efficiency). Thus in most cases this theorem alleviates the need to prove that full discretizations of partial differential equations that satisfy infinite dimensional analogues of (6.1.2) retain the dissipativity of the underlying system; if the full discretization corresponds to applying an algebraically stable Runge-Kutta method to a semi-discretized system which retains the dissipativity of the underlying partial differential equation then the full discretization must also preserve the dissipativity of the underlying system. The task of discretizing dissipative partial differential equations in space so as to produce dissipative semi-discrete systems is far from trivial however, and is beyond the scope of this project.

Finally in Section 6.2 although the numerical solution does not define a discrete

dynamical system on \mathbb{R}^m , Theorem 6.2.3 shows that if the step-size is sufficiently small then it does define a continuous discrete dynamical system in a natural way on its absorbing set which contains the global attractor \mathcal{A}_h of the numerical solution, and that the numerical solution has the same fixed points as the underlying system (6.1.1–2).

In Section 6.3 we consider (6.1.1–2) under the additional assumption that \mathbf{f} is globally Lipschitz. Theorem 6.3.1 shows that for these systems the numerical solution by any Runge-Kutta method with positive weights defines a continuous dissipative discrete dynamical system with a global attractor \mathcal{A}_h if h is sufficiently small. We also present an example of a globally Lipschitz dissipative system not of the form (6.1.1–2) for which the numerical solution by the forward Euler method is not dissipative for any $h > 0$. This implies that Theorem 6.3.1 cannot be extended to arbitrary globally Lipschitz dissipative systems.

In Section 6.4 we consider the numerical approximation of (6.1.1) under the assumptions that $\mathbf{f}: \mathbb{R}^m \rightarrow \mathbb{R}^m$ is locally Lipschitz and

$$\langle \mathbf{f}(\mathbf{y}), \mathbf{y} \rangle < 0 \quad \text{for } \|\mathbf{y}\| > R \quad (6.1.3)$$

for some $R \geq 0$. Recall from Section 2.2.1 that (6.1.1, 6.1.3) does define a dissipative system and that (6.1.3) is a generalization of (6.1.2). Although Theorem 3.6.20 implies that the Runge-Kutta defining equations (3.2.1–2) are always soluble for a certain class of Runge-Kutta methods when \mathbf{f} satisfies (6.1.3), since (6.1.3) is a generalization of (6.1.2) this solution will not, in general be unique, and once again we must consider the dissipativity of the numerical solution in terms of the generalized concept of dissipativity for multi-valued maps from Section 2.5.2. We show that the numerical solution of (6.1.1, 6.1.3) defined by either the one- or two-stage theta method with $\theta \in [1/2, 1]$ is dissipative in this generalized sense for any step-size $h > 0$. Since the two-stage theta method (3.5.2) is A-stable but not algebraically stable for $\theta \in [1/2, 1)$ this shows that algebraic stability is not a necessary condition for the numerical solution to retain the dissipativity of the underlying system.

6.2 Dissipativity of Algebraically Stable Methods

We now consider whether the numerical solution defined by a Runge-Kutta method is dissipative. We begin with an example which shows that a numerical discretization of (6.1.1–2) need not in general inherit the dissipativity of that system.

Example 6.2.1 Consider the class of linear scalar systems (3.4.1) with λ real and negative. Note that this system is dissipative and satisfies (6.1.2) with $\alpha = 0, \beta = -\lambda$. Solving numerically with the forward Euler method we obtain the numerical solution

$$y_n = (1 + h\lambda)^n y_0$$

which is dissipative for $h < 2/(-\lambda)$. If $h > 2/(-\lambda)$ the numerical solution will become unbounded. Thus to ensure numerical dissipativity for linear problems we must impose an upper bound on the step-size when using the forward Euler method.

It is easy to show that every complex contractive problem of the form (3.4.1) can be written as a real linear dissipative system in \mathbb{R}^2 and it follows from this that an upper bound must be placed on the step-size to maintain dissipativity when a non A-stable method is used to solve a linear dissipative system. If a linear dissipative system is approximated numerically using an A-stable Runge-Kutta method then the numerical solution will be dissipative for all $h > 0$.

For nonlinear problems the situation is worse; consider the numerical solution of

$$\frac{dy}{dt} = -y^3, \quad y(0) = y_0 \tag{6.2.1}$$

using the forward Euler method. Note that $\langle f(y), y \rangle \leq 1 - y^2$ so that (6.2.1) defines a dissipative system of the form (6.1.1–2). The numerical solution has the property that if $|y_0| < \sqrt{2/h}$ then $|y_n| \rightarrow 0$ as $n \rightarrow \infty$, whilst if $|y_0| > \sqrt{2/h}$ then $|y_{n+1}| > |y_n|$ and $|y_n| \rightarrow \infty$. Hence the numerical solution defined by the forward Euler method is not dissipative for any $h > 0$. \square

Thus whenever a non A-stable method is used to solve (6.1.1–2), a restriction must be imposed on the step-size used to ensure dissipativity for linear problems. For nonlinear problems there is no obvious analogue of λ , and it is then necessary to impose bounds on h which are initial data dependent. For general nonlinear systems and general non A-stable methods even these initial data dependent bounds can be hard to

derive explicitly, and we will not seek such bounds. Instead, and in order to obtain robust numerical schemes, we will seek methods which generate dissipative numerical solutions for any fixed step-size $h > 0$. Example 6.2.1 then implies that such a method must be A-stable. Initially however we will further restrict our attention and consider the numerical approximation of (6.1.1–2) by algebraically stable Runge-Kutta methods.

We wish to show that the map defined by the numerical solution is dissipative, but as we showed in Example 3.6.19 the Runge-Kutta defining equations (3.2.1–2) may have multiple solutions when applied to (6.1.1–2) and so the numerical solution does not in general define a discrete dynamical system. We could impose additional structure on the problem, such as a one-sided Lipschitz condition, that would ensure the existence of a unique numerical solution, but as we noted in Section 2.2.2 this would exclude many of the problems in which we are interested, and so we do not do this. The approach we will follow is to accept that (3.2.1–2) may have multiple solutions, and hence that the numerical solution defines a multi-valued map. We will now show that, when an algebraically stable Runge-Kutta method is applied to (6.1.1–2), this multi-valued map is dissipative in the generalized sense of Section 2.5.2. The proof of this theorem requires Proposition 3.2.5 and the notation established on page 86.

Theorem 6.2.2 *Suppose (6.1.1–2) is approximated numerically using an algebraically stable Runge-Kutta method. Then for any fixed step-size $h > 0$ the multi-valued map generated by the numerical method is dissipative in the generalized sense of Section 2.5.2 and the open ball $B(0, R)$ is an absorbing set for any $R > \sqrt{\alpha/\beta + hC(0, h)}$ where C is defined in (6.2.7).*

Proof. First suppose that the method is DJ-irreducible, then algebraic stability and (3.2.22) implies

$$\|\mathbf{y}_{n+1}\|^2 \leq \|\mathbf{y}_n\|^2 + 2h \sum_{i=1}^s b_i \langle \mathbf{Y}_i, \mathbf{f}(\mathbf{Y}_i) \rangle. \quad (6.2.2)$$

Now since (6.1.2) holds it follows that

$$\begin{aligned} \|\mathbf{y}_{n+1}\|^2 &\leq \|\mathbf{y}_n\|^2 + 2h \sum_{i=1}^s b_i [\alpha - \beta \|\mathbf{Y}_i\|^2] \\ &= \|\mathbf{y}_n\|^2 + 2h [\alpha - \beta \|\mathbf{Y}\|_B^2]. \end{aligned} \quad (6.2.3)$$

Hence given any $\varepsilon > 0$ it follows that either

$$\|y_{n+1}\|^2 \leq \|y_n\|^2 - 2h\beta\varepsilon \quad (6.2.4)$$

or

$$\|Y\|_B^2 \leq \frac{\alpha}{\beta} + \varepsilon \quad (6.2.5)$$

where $\|\bullet\|_B$ is defined by (3.6.11). But if (6.2.5) holds then (3.2.23) implies that

$$\|y_{n+1}\|^2 \leq \frac{\alpha}{\beta} + \varepsilon + 2hY^T(BE \otimes I_m)F(Y) + h^2\|(E \otimes I_m)F(Y)\|_B^2 \quad (6.2.6)$$

where we have used the notation defined on page 86. Now let

$$C(\varepsilon, h) = \sup_{\|X\|_B^2 \leq \alpha/\beta + \varepsilon} \left[2X^T(BE \otimes I_m)F(X) + h\|(E \otimes I_m)F(X)\|_B^2 \right]. \quad (6.2.7)$$

Note that B is positive definite so that $\|\bullet\|_B$ defines a norm and hence the supremum is taken over a compact set. Lipschitz continuity of F follows from Lipschitz continuity of f and hence it follows that $\|F\|$ is uniformly bounded on the set $\{X: \|X\|_B^2 \leq \alpha/\beta + \varepsilon\}$. It now follows that $C(\varepsilon, h)$ is nonnegative and finite for any nonnegative ε, h . C is clearly continuous and increasing in h , and it follows from the continuity of F that C is also continuous and increasing in ε . Now (6.2.6) implies that

$$\|y_{n+1}\|^2 \leq \frac{\alpha}{\beta} + \varepsilon + hC(\varepsilon, h). \quad (6.2.8)$$

Hence either (6.2.4) or (6.2.8) holds at each step and it follows trivially that the multi-valued map generated by the numerical method is dissipative in the generalized sense of Section 2.5.2 and that $B(0, \sqrt{\alpha/\beta + \varepsilon + hC(\varepsilon, h)})$ is an absorbing set. Since ε is arbitrary the result follows for DJ-irreducible methods.

Any solution sequence $\{y_n\}_{n=0}^\infty$ of a DJ-reducible algebraically stable method also defines a solution sequence of the equivalent DJ-irreducible algebraically stable method and hence must enter the ball $B(0, \sqrt{\alpha/\beta + \varepsilon + hC(\varepsilon, h)})$ (where C is defined in terms of the reduced method). If the Runge-Kutta defining equations (3.2.1–2) are not soluble for some y_n then $G_h(y_n) = \emptyset$ where $G_h(\bullet)$ is the generalized evolution map and since the empty set is contained in any set, it follows that the multi-valued map defined by a DJ-reducible method is dissipative in the generalized sense of Section 3.6, regardless

of whether the defining equations are soluble. ■

Remarks (i) Theorem 3.6.18 ensures that there exists a (not necessarily unique) solution sequence $\{\mathbf{y}_n\}_{n=0}^{\infty}$ for any DJ-irreducible algebraically stable Runge-Kutta method (3.2.1–2), with invertible A , applied to (6.1.1–2). Then Theorem 6.2.2 implies that any such solution sequence must enter and then remain inside the absorbing set for n sufficiently large.

(ii) For a DJ-reducible method there need not exist a solution sequence, but if there does then it must also enter the absorbing set. Hence a numerical simulation may fail, or it may enter the absorbing set, but it cannot blow up. Moreover Theorem 3.6.4 can be used to imply that for any initial condition \mathbf{y}_0 there exists $h(\mathbf{y}_0) > 0$ such that for $h \in (0, h(\mathbf{y}_0))$ there does exist a solution sequence $\{\mathbf{y}_n\}_{n=0}^{\infty}$ for the method.

(iii) Proposition 3.6.17 together with Theorem 3.6.18 imply the existence of an upper bound on $\|\mathbf{Y}\|_B^2$ for the solution of (3.2.1) at each step if the method is DJ-irreducible. From the proof of Theorem 6.2.2 we can write down such a bound explicitly. Since $\|\mathbf{y}_{n+1}\|^2 \geq 0$, it follows from (6.2.3) that any solution of (3.2.1) satisfies

$$\|\mathbf{Y}\|_B^2 \leq \frac{\alpha}{\beta} + \frac{1}{2h\beta} \|\mathbf{y}_n\|^2. \quad (6.2.9)$$

Note that this bound also holds for the non-redundant stages of a DJ-reducible methods, where in that case the norm $\|\bullet\|_B$ in (6.2.9) is the norm induced by the equivalent DJ-irreducible method.

(iv) From the proof of Theorem 6.2.2 by setting $\varepsilon = 0$ we deduce that for any $h > 0$ the ball

$$B(0, \sqrt{\alpha/\beta + hC(0, h)}), \quad (6.2.10)$$

where C is defined by (6.2.7), is forward invariant for the numerical method. By this we mean that if $\mathbf{y}_n \in B(0, \sqrt{\alpha/\beta + hC(0, h)})$ then $\mathbf{y}_{n+1} \in B(0, \sqrt{\alpha/\beta + hC(0, h)})$. Note however that the corresponding stage values \mathbf{Y}_i need not be contained in the forward invariant set.

(v) Notice that $hC(0, h) \rightarrow 0$ as $h \rightarrow 0$ hence given any $\varepsilon > 0$ there exists $H(\varepsilon) > 0$ such that for $h < H(\varepsilon)$ the ball $B(0, \sqrt{\alpha/\beta} + \varepsilon)$ is an absorbing set.

We would like to combine the local uniqueness result Proposition 3.6.3 with the a priori bound (6.1.2) on the solution of (3.2.1) to prove global uniqueness of the solution

of (3.2.1–2) when $y_n \in B$ where B is some bounded neighbourhood of the absorbing set. However the nature of the bound given by (6.2.9) does not allow us to do this. For if we fix $h > 0$ then (6.2.9) defines a set in which all solutions of (3.2.1) must lie. To ensure that there is a unique such solution we must also ensure that (3.6.7) is satisfied, but in general we cannot do this, since reducing h to satisfy (3.6.7) will enlarge the set defined by (6.2.9), which will in turn require a smaller h to satisfy (3.6.7) and so on.

Although we cannot derive a global uniqueness result, the local existence and uniqueness result Proposition 3.6.3 enables us, via Theorem 3.6.4, to prove that the numerical method defines a continuous discrete dynamical system on the absorbing set for h sufficiently small. Using the results of Chapter 4 we also show that the numerical solution admits no spurious fixed points for h sufficiently small, even though f is not necessarily globally Lipschitz.

Theorem 6.2.3 *If (6.1.1–2) is approximated numerically using an algebraically stable Runge-Kutta method then for any $B = B(0, R)$, where $R > \sqrt{\alpha/\beta}$, and any neighbourhood $N = \mathcal{N}(B, \varepsilon)$ of B there exists $H = H(B, N) > 0$ such that for $h \in (0, H)$, B is an absorbing set for the numerical solution. Moreover if for $y_n \in B$ the locally unique solution of (3.2.1) defined by Proposition 3.6.3 is used then the numerical solution defines a continuous discrete dynamical system on B so that if $y_0 \in B$ then $y_n \in B$ for all $n \geq 0$, and, furthermore, the stage values $Y_i \in N$ at each step. This implies that the numerical solution admits no spurious fixed points, and possesses a global attractor $\mathcal{A}_h \subset B$.*

Proof. Given a set B as above then by Theorem 6.2.2 B is absorbing for $h < H_1(R)$ for some $H_1(R) > 0$. Now note that by Theorem 6.2.2, Remark (iii), if $y_n \in B$ then $y_{n+1} \in B$ and it follows from Theorem 3.6.4 that for $h < H_2(B, N)$ where $H_2(B, N)$ is defined by (3.6.7) the numerical solution defines a continuous discrete dynamical system on B as required. The dissipativity of the numerical solution implies the existence of a global attractor contained in the absorbing set B , and also that all the fixed points of the numerical solution are contained in B . Now if $h < H_3(B, N)$ where $H_3(B, N)$ is defined by (4.2.11) then Lemma 4.2.5 implies that the numerical solution admits no spurious fixed points within B , and hence the theorem holds with $H = \min(H_1, H_2, H_3)$.

■

In [36] the convergence of the numerical attractor \mathcal{A}_h to the global attractor \mathcal{A} of

the underlying system (6.1.1–2) is considered and it is shown that $\text{dist}(\mathcal{A}_h, \mathcal{A}) \rightarrow 0$ as $h \rightarrow 0$ where $\text{dist}(\bullet, \bullet)$ is as defined in Definition 2.3.5. We do not reproduce that result here, since it is a special case of Theorem 7.2.2; and we will study the behaviour of $\text{dist}(\mathcal{A}_h, \mathcal{A})$ as $h \rightarrow 0$ in Section 7.2.

6.3 Dissipativity Under Global Lipschitz Condition

In this section we will consider the numerical approximation of (6.1.1–2) under the additional assumption that f is globally Lipschitz. Recall (from the remark after Theorem 2.2.2) that under (6.1.2) the system decays at least linearly from infinity. If f is globally Lipschitz then the decay from infinity must also be at most linear so that for $\|y\|$ large

$$-L\|y(t)\| \leq \frac{d}{dt}\|y(t)\| \leq -(\beta - \varepsilon)\|y(t)\|$$

(for some $\varepsilon \geq 0$). We now show that the numerical approximation to such a system by a Runge-Kutta method with positive weights defines a dissipative discrete dynamical system on \mathbb{R}^m .

Theorem 6.3.1 *If f is globally Lipschitz with Lipschitz constant L and (6.1.1–2) is approximated numerically by a Runge-Kutta method with $b_i > 0$ for all i then*

(i) *if*

$$h < H_1 = \frac{2\beta}{\rho^2 L^2 \mathbb{M}} \quad (6.3.1)$$

where $\rho = \max_i 1/b_i$ and $\mathbb{M} = \sum_{i,j=1}^s |m_{ij}|$ then the numerical solution is dissipative in the generalized sense of Section 2.5.2;

(ii) *if*

$$h < H_2 = \frac{1}{La} \quad (6.3.2)$$

then the numerical solution defines a dynamical system on \mathbb{R}^m ;

(iii) *for $h < \min(H_1, H_2)$ the numerical solution defines a continuous dissipative discrete dynamical system on \mathbb{R}^m and possesses a global attractor \mathcal{A}_h . Moreover given any $\varepsilon > 0$ there exists $H > 0$ such that if $h < H$ then $\mathcal{A}_h \in B(0, \sqrt{\alpha/\beta + \varepsilon + HC(\varepsilon, H)})$ where C is defined by (6.2.7).*

Proof. To show dissipativity choose $\varepsilon > 0$ and let $k = 1 + (\varepsilon\beta/2\alpha)$ so that

$$\frac{k\alpha}{\beta} = \frac{\alpha}{\beta} + \frac{\varepsilon}{2}.$$

Recall Proposition 3.2.5 and apply (6.1.2) to obtain

$$\begin{aligned}
 \|\mathbf{y}_{n+1}\|^2 &\leq \|\mathbf{y}_n\|^2 + 2h \sum_{i=1}^s b_i [\alpha - \beta \|\mathbf{Y}_i\|^2] - h^2 \sum_{i,j=1}^s m_{ij} \langle \mathbf{f}(\mathbf{Y}_i), \mathbf{f}(\mathbf{Y}_j) \rangle \\
 &\leq \|\mathbf{y}_n\|^2 + 2h \left[\alpha - \beta \|\mathbf{Y}\|_B^2 \right] + h^2 \left\| \sum_{i,j=1}^s m_{ij} \langle \mathbf{f}(\mathbf{Y}_i), \mathbf{f}(\mathbf{Y}_j) \rangle \right\| \\
 &= \|\mathbf{y}_n\|^2 + 2h \left[\alpha - \frac{\beta}{k} \|\mathbf{Y}\|_B^2 \right] - 2h\beta \left(1 - \frac{1}{k}\right) \|\mathbf{Y}\|_B^2 \\
 &\quad + h^2 \left\| \sum_{i,j=1}^s m_{ij} \langle \mathbf{f}(\mathbf{Y}_i), \mathbf{f}(\mathbf{Y}_j) \rangle \right\|
 \end{aligned} \tag{6.3.3}$$

By Result 2.6.2 the underlying system has a fixed point \mathbf{x}^* such that $\|\mathbf{x}^*\| \leq \sqrt{\alpha/\beta}$. Thus by Lipschitz continuity $\|\mathbf{f}(\mathbf{0})\| \leq L\sqrt{\alpha/\beta}$ and hence letting $c = L\sqrt{\alpha/\beta}$ implies that

$$\begin{aligned}
 \|\mathbf{f}(\mathbf{Y}_i)\| &\leq L\|\mathbf{Y}_i\| + c \\
 &\leq \rho b_i L \|\mathbf{Y}_i\| + c \\
 &\leq \rho L \|\mathbf{Y}\|_B + c.
 \end{aligned}$$

Thus

$$\begin{aligned}
 |\langle \mathbf{f}(\mathbf{Y}_i), \mathbf{f}(\mathbf{Y}_j) \rangle| &\leq \|\mathbf{f}(\mathbf{Y}_i)\| \|\mathbf{f}(\mathbf{Y}_j)\| \\
 &\leq (\rho L \|\mathbf{Y}\|_B + c)^2,
 \end{aligned}$$

and (6.3.3) implies that

$$\|\mathbf{y}_{n+1}\|^2 \leq \|\mathbf{y}_n\|^2 + 2h \left[\alpha - \frac{\beta}{k} \|\mathbf{Y}\|_B^2 \right] - 2h\beta \left(1 - \frac{1}{k}\right) \|\mathbf{Y}\|_B^2 + h^2 \mathbf{M} (\rho L \|\mathbf{Y}\|_B + c)^2. \tag{6.3.4}$$

Now assuming $\mathbf{M} \neq 0$ (otherwise the method is irreducible and algebraically stable and the previous theory applies) let

$$H = \min_{\|\mathbf{x}\|_B^2 \geq \alpha/\beta + \varepsilon} \frac{2\beta(1 - \frac{1}{k})\|\mathbf{x}\|_B^2}{\mathbf{M}(\rho L \|\mathbf{x}\|_B + c)^2}$$

and notice that the minimum is achieved with $\|X\|_B^2 = \alpha/\beta + \varepsilon$ and that H is strictly positive. Suppose $h < H$ then by (6.3.4) and definition of H either

$$\|y_{n+1}\|^2 \leq \|y_n\|^2 + 2h \left[\alpha - \frac{\beta}{k} \|Y\|_B^2 \right] \quad (6.3.5)$$

or

$$\|Y\|_B^2 \leq \frac{\alpha}{\beta} + \varepsilon. \quad (6.3.6)$$

Now (6.3.5) implies that either

$$\|y_{n+1}\|^2 \leq \|y_n\|^2 - h\beta\varepsilon/k \quad (6.3.7)$$

or

$$\begin{aligned} \|Y\|_B^2 &\leq \frac{k\alpha}{\beta} + \frac{\varepsilon}{2} \\ &= \frac{\alpha}{\beta} + \varepsilon. \end{aligned}$$

Hence we have deduced that for $h < H$ either (6.3.6) or (6.3.7) holds at each step. Now follow the proof of Theorem 6.2.2 to deduce that (6.3.6) implies (6.2.8) holds and hence that the numerical solution is dissipative and that $B(0, \sqrt{\alpha/\beta + \varepsilon + hC(\varepsilon, h)})$ is an absorbing set.

Finally, notice that as $\|X\|_B \rightarrow \infty$

$$\frac{2\beta(1 - \frac{1}{k})\|X\|_B^2}{\mathbf{M}(\rho L\|X\|_B + c)^2} \rightarrow \frac{2\beta(1 - \frac{1}{k})}{\rho^2 L^2 \mathbf{M}}$$

and as $\varepsilon \rightarrow \infty$

$$\frac{2\beta(1 - \frac{1}{k})}{\rho^2 L^2 \mathbf{M}} \rightarrow H_1$$

hence, given any $h < H_1$, for ε sufficiently large $h < H$, and the numerical solution is dissipative.

By Result 3.6.1 if (6.3.2) holds the numerical method defines a dynamical system on \mathbb{R}^m and continuity may be established as in the proof of Theorem 3.6.4. From (i) and (ii) for $h < \min(H_1, H_2)$ the numerical solution defines a dissipative dynamical system and possesses a global attractor \mathcal{A}_h . Since C is an increasing function in h it follows from above that $B(0, \sqrt{\alpha/\beta + \varepsilon + HC(\varepsilon, H)})$ is an absorbing set for any $h < H$

and that the global attractor \mathcal{A}_h of the numerical solution is contained in this set. ■

It may not seem surprising that an arbitrary method (with positive weights) will preserve the dissipativity of (6.1.1–2) when \mathbf{f} is globally Lipschitz. Indeed we might hope that the numerical solution will preserve the dissipativity of the underlying problem when \mathbf{f} is globally Lipschitz for any dissipative dynamical system, not just those defined by (6.1.1–2). However the following example shows that there exist dissipative systems with \mathbf{f} globally Lipschitz for which the numerical solution defined by the forward Euler method is not dissipative for any $h > 0$, and hence that Theorem 6.3.1 cannot be extended to cover general dissipative systems where \mathbf{f} is globally Lipschitz.

Example 6.3.2 Consider the two-dimensional system in polar coordinates defined by

$$\dot{r} = -r^{k-1} \quad \dot{\theta} = 1 \quad (6.3.8)$$

for some $k \in (1, 2)$. Converting to Cartesian coordinates let $\mathbf{u} = (x, y)^T$ then, noting that $r = \|\mathbf{u}\|$, we have that

$$\mathbf{f}(\mathbf{u}) = \left(-x\|\mathbf{u}\|^{k-2} - y, -y\|\mathbf{u}\|^{k-2} + x \right)^T$$

and hence

$$\begin{aligned} \langle \mathbf{f}(\mathbf{u}), \mathbf{u} \rangle &= -\|\mathbf{u}\|^k \\ &= \alpha - \beta\|\mathbf{u}\|^k \end{aligned}$$

for all $\mathbf{u} \in \mathbb{R}^m$ where $\alpha = 0$ and $\beta = 1$. Note that \mathbf{f} satisfies (2.2.9) with $R = 0$ so that by Theorem 2.2.4 this is a dissipative system, and moreover since $R = 0$ it follows that $\|\mathbf{u}(t)\| \rightarrow 0$ as $t \rightarrow \infty$ for any initial condition \mathbf{u}_0 .

Now consider the numerical approximation of this system using the forward Euler method. Let $\mathbf{u}_n = (x_n, y_n)^T$ then

$$\begin{aligned} x_{n+1} &= x_n - h(x_n\|\mathbf{u}_n\|^{k-2} - y_n) \\ y_{n+1} &= y_n - h(y_n\|\mathbf{u}_n\|^{k-2} - x_n). \end{aligned}$$

Thus

$$\|\mathbf{u}_{n+1}\|^2 = \|\mathbf{u}_n\|^2 - 2h\|\mathbf{u}_n\|^k + h^2(\|\mathbf{u}_n\|^2 + \|\mathbf{u}_n\|^{2(k-1)}). \quad (6.3.9)$$

Now note that if $\|\mathbf{u}_n\|^{2-k} \geq 2/h$ then

$$h^2\|\mathbf{u}_n\|^2 - 2h\|\mathbf{u}_n\|^k \geq 0$$

and hence by (6.3.9), if $\|\mathbf{u}_n\|^{2-k} \geq 2/h$ then

$$\|\mathbf{u}_{n+1}\|^2 \geq \|\mathbf{u}_n\|^2 + h^2\|\mathbf{u}_n\|^{2(k-1)}.$$

Thus for any $h > 0$, choose \mathbf{u}_0 such that $\|\mathbf{u}_0\| \geq (2/h)^{\frac{1}{2-k}}$ and then, by induction, $\|\mathbf{u}_n\| \rightarrow \infty$ as $n \rightarrow \infty$ and hence the numerical solution cannot be dissipative for any $h > 0$.

Note that \mathbf{f} as defined above is Lipschitz on any open set that does not contain the origin, but that \mathbf{f} is not Lipschitz at the origin. If we modify (6.3.8) so that $\dot{r} = -r$ for $r \leq 1$ then $\mathbf{f}(\mathbf{u}) = (-x - y, -y + x)^T$ for $\|\mathbf{u}\| \leq 1$ and the dynamics of the numerical solution are not affected for $\|\mathbf{u}_0\| \geq 1$ and \mathbf{f} becomes globally Lipschitz. \square

Theorem 6.3.1 showed that an arbitrary Runge-Kutta method with positive weights preserves the dissipativity of (6.1.1-2) if \mathbf{f} is globally Lipschitz. But if we relax condition (6.1.2) and suppose that

$$\langle \mathbf{f}(\mathbf{u}), \mathbf{u} \rangle \leq \alpha - \beta\|\mathbf{u}\|^k$$

for any $k < 2$ then Example 6.3.2 shows that the numerical solution need not preserve the dissipativity of the underlying system and hence that Theorem 6.3.1 is optimal in this sense.

Note from (6.3.9) that since $k \leq 2$

$$\|\mathbf{u}_{n+1}\|^2 \leq \|\mathbf{u}_n\|^2 - 2h\|\mathbf{u}_n\|(1 - h\|\mathbf{u}_n\|).$$

Hence if $\|\mathbf{u}_0\| < 1/h$ then $\|\mathbf{u}_n\| \rightarrow 0$ as $n \rightarrow \infty$ and the origin (which is the global attractor of the underlying system) is a local attractor for the numerical solution. In Chapter 7 we will show that if any dissipative dynamical system with \mathbf{f} locally Lipschitz is approximated numerically using a Runge-Kutta method then although the numerical solution need not be dissipative for h sufficiently small it will possess a local attractor \mathcal{A}_h in a neighbourhood of the global attractor \mathcal{A} of the underlying system.

6.4 Dissipativity of Theta Methods

In this section we will consider the dissipative system (6.1.1,6.1.3) and will show that the dissipativity of the system is preserved under numerical approximation by both the one- and two-stage theta methods with $\theta \in [1/2, 1]$ and any $h > 0$. We first prove the result for the one-stage theta method (3.5.1).

Theorem 6.4.1 *Suppose $f: \mathbb{R}^m \rightarrow \mathbb{R}^m$ is locally Lipschitz and (6.1.1,6.1.3) is approximated numerically using the one-stage theta method (3.5.1) with $\theta \in [1/2, 1]$. Then for any fixed step-size $h > 0$ the multi-valued map generated by the numerical solution is dissipative in the generalized sense of Section 2.5.2 and the closed ball*

$$\overline{B}(0, R + h(1 - \theta)M)$$

where

$$M = \sup_{\mathbf{x} \in \overline{B}(0, R)} \|f(\mathbf{x})\|$$

is an absorbing set for $\theta \in [1/2, 1)$, and any open set containing $\overline{B}(0, R)$ is absorbing for $\theta = 1$.

Proof. Note that the method is algebraically stable for $\theta \in [1/2, 1]$ and hence by (3.2.22)

$$\|\mathbf{y}_{n+1}\|^2 \leq \|\mathbf{y}_n\|^2 + 2h\langle \mathbf{Y}_1, f(\mathbf{Y}_1) \rangle \quad (6.4.1)$$

and (6.1.3) implies that $\|\mathbf{y}_{n+1}\| < \|\mathbf{y}_n\|$ unless $\|\mathbf{Y}_1\| \leq R$. But note that if $\|\mathbf{Y}_1\| \leq R$ then

$$\begin{aligned} \mathbf{Y}_1 &= \mathbf{y}_n + h\theta f(\mathbf{Y}_1), \\ \mathbf{y}_{n+1} &= \mathbf{y}_n + hf(\mathbf{Y}_1) \end{aligned}$$

hence

$$\mathbf{y}_{n+1} = \mathbf{Y}_1 + h(1 - \theta)f(\mathbf{Y}_1)$$

and

$$\|\mathbf{y}_{n+1}\| \leq R + h(1 - \theta)M.$$

Thus if $\|\mathbf{y}_0\| \leq R + h(1 - \theta)M$ then $\|\mathbf{y}_n\| \leq R + h(1 - \theta)M$ for all $n \geq 0$ and hence $\overline{B}(0, r)$ is a forward invariant set for the numerical solution for any $r \geq R + h(1 - \theta)M$.

To complete the proof we must show that for any \mathbf{y}_0 there exists $n^*(\mathbf{y}_0)$ such that for all $n \geq n^*$ $\mathbf{y}_n \in \overline{B}(0, r)$. Now, since $\overline{B}(0, r)$ is forward invariant it is sufficient to show that $\mathbf{y}_n \in \overline{B}(0, r)$ for some n . We will prove this by contradiction. Let $X = \{\mathbf{x} : R + \delta \leq \|\mathbf{x}\| \leq \|\mathbf{y}_0\|\}$ for some $\delta \in (0, \|\mathbf{y}_0\| - R)$. Then X is compact and since $\langle \mathbf{x}, \mathbf{f}(\mathbf{x}) \rangle < 0$ on X there exists $\varepsilon > 0$ such that $\langle \mathbf{x}, \mathbf{f}(\mathbf{x}) \rangle \leq -\varepsilon$ for all $\mathbf{x} \in X$. Now suppose $\|\mathbf{y}_n\| \geq R + \delta$ for all n . Notice that

$$\mathbf{Y}_1 = \theta \mathbf{y}_{n+1} + (1 - \theta) \mathbf{y}_n.$$

Therefore $\|\mathbf{Y}_1\| \geq R + \delta$ for all n . Thus by (6.4.1)

$$\|\mathbf{y}_{n+1}\|^2 \leq \|\mathbf{y}_n\|^2 - 2h\varepsilon$$

and

$$\|\mathbf{y}_n\|^2 \leq \|\mathbf{y}_0\|^2 - 2nh\varepsilon$$

which leads to a contradiction for n sufficiently large. For $\theta \in [1/2, 1)$ the result follows on choosing $\delta \in (0, h(1 - \theta)M)$, and given any open set U containing $\overline{B}(0, R)$ the result follows for $\theta = 1$ by choosing δ sufficiently small such that $\overline{B}(0, R + \delta) \subset U$. ■

Example 6.2.1 showed that A-stability is a necessary condition for the numerical solution to preserve the dissipativity of the underlying system, and we showed in Theorem 6.2.2 that, at least under condition (6.1.2), algebraic stability is sufficient. For the final result in this chapter we will show that algebraic stability is not a necessary condition for the numerical solution to preserve the dissipativity of the underlying system. The two-stage theta method is not algebraically stable for $\theta < 1$, but it is A-stable for all $\theta \in [1/2, 1]$. We will show that if (6.1.1, 6.1.3) is approximated numerically by the two-stage theta method with $\theta \in [1/2, 1]$ then the dissipativity of the underlying system is preserved for any $h > 0$. The result follows as a corollary to Theorem 6.4.1 using the relationship between solution sequences for the one and two-stage theta methods.

Corollary 6.4.2 *Suppose $\mathbf{f}: \mathbb{R}^m \rightarrow \mathbb{R}^m$ is locally Lipschitz and (6.1.1, 6.1.3) is approximated numerically using the two-stage theta method (3.5.2) with $\theta \in [1/2, 1]$. Then for any fixed step-size $h > 0$ the multi-valued map generated by the numerical solution*

is dissipative in the generalized sense of Section 2.5.2 and the closed ball

$$\overline{B}(0, R + h(1 - \theta)M)$$

where

$$M = \sup_{\mathbf{x} \in \overline{B}(0, R)} \|f(\mathbf{x})\|$$

is an absorbing set for $\theta \in [1/2, 1)$, and any open set containing $\overline{B}(0, R)$ is absorbing for $\theta = 1$.

Proof. Suppose $\{\mathbf{y}_n\}_{n=0}^{\infty}$ is a solution sequence for the two-stage theta method (3.5.2) with $\theta \in [1/2, 1]$. Then from Example 3.5.1 there exists $\{\mathbf{x}_n\}_{n=0}^{\infty}$ which defines a solution sequence for the one-stage theta method (3.5.1) and satisfies

$$\mathbf{y}_n = (1 - \theta)\mathbf{x}_n + \theta\mathbf{x}_{n+1}. \quad (6.4.2)$$

Now suppose that $\theta \in [1/2, 1)$ then the one-stage theta method defines a dissipative numerical solution to (6.1.1, 6.1.3) with $\overline{B}(0, R + h(1 - \theta)M)$ as an absorbing set. Hence there exists $n^*(\mathbf{x}_0)$ such that $\mathbf{x}_n \in \overline{B}(0, R + h(1 - \theta)M)$ for all $n \geq n^*(\mathbf{x}_0)$. Hence by (6.4.2)

$$\mathbf{y}_n \in \overline{B}(0, R + h(1 - \theta)M) \quad (6.4.3)$$

for all $n \geq n^*(\mathbf{x}_0) + 1$, and upon noting that \mathbf{x}_0 is uniquely determined by \mathbf{y}_0 it follows that (6.4.3) holds for all $n \geq \tilde{n}^*(\mathbf{y}_0)$ where $\tilde{n}^*(\mathbf{y}_0) = n^*(\mathbf{x}_0) + 1$ and hence the numerical solution is dissipative for $\theta \in [1/2, 1)$ with the relevant absorbing set. The result follows in a similar manner for the case $\theta = 1$. ■

Chapter 7

Attractors and Invariant Sets

7.1 Introduction

In this chapter we will consider the numerical approximation of attractors and invariant sets of dynamical systems by Runge-Kutta methods. If the underlying system possesses a local or global attractor we show that for h sufficiently small the numerical solution possesses an attractor and that this attractor “converges” to the attractor of the underlying system as $h \rightarrow 0$. We will also show that if the numerical solution possesses a continuous (in h) branch of uniformly bounded invariant sets then these “converge” to an invariant set of the underlying system as $h \rightarrow 0$.

Specifically we consider dynamical systems generated by

$$\frac{d\mathbf{y}}{dt} = \mathbf{f}(\mathbf{y}) \tag{7.1.1}$$

for $t \in [0, \infty)$, $\mathbf{f}: \mathbb{R}^m \rightarrow \mathbb{R}^m$, $\mathbf{y}(t) \in \mathbb{R}^m$ and an arbitrary initial condition $\mathbf{y}(0) = \mathbf{y}_0$. On an invariant set every point has negative orbit, and in that case we can consider $t \in (-\infty, \infty)$. Throughout this chapter we will assume that \mathbf{f} is at least locally Lipschitz. Unlike in the previous two chapters, we will not impose any structural assumptions on \mathbf{f} in this chapter.

In Sections 7.2 and 7.3 we also assume that this dynamical system possesses a local or global attractor \mathcal{A} . Proposition 7.2.1 shows that for h sufficiently small the numerical solution defined by any Runge-Kutta method (3.2.1–2) defines a discrete dynamical system on a neighbourhood of \mathcal{A} and possesses a local attractor \mathcal{A}_h contained in this neighbourhood. Having established this, the really interesting question is:

Does \mathcal{A}_h converge to \mathcal{A} as $h \rightarrow 0$, and if so, in what sense does this convergence occur ?

Since \mathcal{A} and \mathcal{A}_h are both compact subsets of \mathbb{R}^m the natural concept of convergence to use is convergence as sets in the Hausdorff metric, as defined in Definition 2.3.5, and we will try to prove that

$$\text{dist}_H(\mathcal{A}, \mathcal{A}_h) \rightarrow 0 \quad \text{as} \quad h \rightarrow 0.$$

Since the Hausdorff distance between \mathcal{A} and \mathcal{A}_h is defined to be the maximum of the two semi-distances between \mathcal{A} and \mathcal{A}_h , to establish convergence in the Hausdorff metric we will need to show that both semi-distances tend to zero as $h \rightarrow 0$.

In Theorem 7.2.2 we show that

$$\text{dist}(\mathcal{A}_h, \mathcal{A}) \rightarrow 0 \quad \text{as} \quad h \rightarrow 0. \quad (7.1.2)$$

When this holds the numerical approximation is said to be *upper semicontinuous at $h = 0$* . Equation (7.1.2) implies that given any $\varepsilon > 0$ there exists $h_0 > 0$ such that if $h \in (0, h_0)$ then $\mathcal{A}_h \subseteq \mathcal{N}(\mathcal{A}, \varepsilon)$, and hence, roughly speaking, for h small every point on the numerical attractor \mathcal{A}_h must be close to a point of \mathcal{A} ; but not visa versa – there may be points of \mathcal{A} not approximated by the numerical attractor.

Note that in general \mathcal{A}_h will only be a local attractor, even when the attractor \mathcal{A} that it is approximating is globally attracting. However, as we saw in Chapter 6, if we impose additional conditions, such as requiring that the dynamical system is of the form (6.1.1–2) and that the Runge-Kutta method used is algebraically stable, then we can ensure that the numerical attractor \mathcal{A}_h is globally attracting.

The proof of (7.1.2) follows the method of Hale, Lin & Raugel [29] and Temam [58]. In both of those works upper semicontinuity was proved for certain perturbations $S_\lambda(t)$ of an infinite dimensional evolution operator $S_{\lambda_0}(t)$. As well as straightforward perturbations of the infinite dimensional system the theory in [29, 58] covers the case where $S_\lambda(t)$ represents certain finite dimensional spatial discretizations of both parabolic and hyperbolic partial differential equations. Hale, Lin & Raugel [29] also considered the case of one-step temporal discretizations (without spatial discretization), and this has since been extended to multistep methods by Hill and Süli [33].

In [36] we derived (7.1.2) in the case where \mathcal{A} is the global attractor of a dissipative

dynamical system of the form (6.1.1–2), and the perturbation corresponds to temporal discretization with a Runge-Kutta method. Theorem 7.2.2 extends this result to any attractor of any dynamical system (7.1.1) (for which \mathbf{f} is locally Lipschitz).

A related result can be found in Kloeden and Lorenz [42]. There it is shown that if $\mathbf{f} \in \mathcal{C}^1(U, \mathbb{R}^m)$ and its derivatives are uniformly bounded for a dissipative dynamical system, and the system is approximated numerically using a one-step method which satisfies a suitable uniform local error bound then for any uniformly asymptotically stable (u.a.s.) set Λ of the underlying system the numerical solution possesses a u.a.s. set Λ_h for h sufficiently small which satisfies

$$\text{dist}_H(\Lambda_h, \Lambda) \rightarrow 0 \quad \text{as} \quad h \rightarrow 0. \quad (7.1.3)$$

Since a local or global attractor \mathcal{A} is u.a.s. we can apply this result with $\Lambda = \mathcal{A}$. The uniform asymptotic stability of Λ_h implies that it attracts a neighbourhood of itself, and hence that it contains a local attractor $\mathcal{A}_h = \omega_h(\Lambda_h) \subseteq \Lambda_h$ and then

$$\text{dist}(\mathcal{A}_h, \mathcal{A}) \leq \text{dist}(\Lambda_h, \mathcal{A}) = \text{dist}(\Lambda_h, \Lambda) \leq \text{dist}_H(\Lambda_h, \Lambda)$$

and upper semicontinuity follows from (7.1.3). The main difference between Theorem 7.2.2 and the result of [42] is that Kloeden and Lorenz make quite strong continuity and differentiability conditions on \mathbf{f} to ensure that the numerical solution defines a discrete dynamical system on \mathbb{R}^m and satisfies a suitable uniform local error bound; whereas we only assume that \mathbf{f} is locally Lipschitz and actually prove that the numerical solution defines a discrete dynamical system on a neighbourhood of \mathcal{A} , and we have already derived a uniform global error bound for when \mathbf{f} is locally Lipschitz in Proposition 3.6.8.

Recall from Section 2.3 that we defined $W(\mathcal{E}^*)$ to be the union of the unstable manifolds of a set of hyperbolic fixed points. In Section 7.3 we consider the numerical approximation of \mathcal{A} under the assumption that $\mathcal{A} = \overline{W(\mathcal{E}^*)}$; that is, \mathcal{A} is equal to the closure of the union of the unstable manifolds of its hyperbolic fixed points. We also assume that \mathbf{f} is \mathcal{C}^1 on a neighbourhood of \mathcal{E}^* , the set of hyperbolic fixed points of the attractor. In Proposition 7.3.2 we show that

$$\text{dist}(\overline{W(\mathcal{E}^*)}, \overline{W_h(\mathcal{E}^*)}) \rightarrow 0 \quad \text{as} \quad h \rightarrow 0. \quad (7.1.4)$$

Since $\overline{W_h(\mathcal{E}^*)}$, the closure of the union of the unstable manifolds of the hyperbolic fixed points for the discretized problem, is a subset of the numerical attractor \mathcal{A} and we are working under the assumption that $\mathcal{A} = \overline{W(\mathcal{E}^*)}$, equation (7.1.4) implies that

$$\text{dist}(\mathcal{A}, \mathcal{A}_h) \rightarrow 0 \quad \text{as} \quad h \rightarrow 0. \quad (7.1.5)$$

When this holds the numerical approximation is said to be *lower semicontinuous at* $h = 0$.

Theorem 7.3.3 combines (7.1.2) and (7.1.5) to show that if $\mathcal{A} = \overline{W(\mathcal{E}^*)}$ then \mathcal{A}_h converges to \mathcal{A} in the Hausdorff metric; or, roughly speaking, for h sufficiently small every point of \mathcal{A}_h is close to a point of \mathcal{A} , and there is a point of \mathcal{A}_h close to every point of \mathcal{A} .

Lower semicontinuity at $h = 0$ is much harder to establish than upper semicontinuity. Hale and Raugel [30] have established lower semicontinuity results for certain perturbations of gradient systems on a Banach spaces with hyperbolic equilibria. As well as the case where the perturbed evolution operator $S_\lambda(t)$ varies continuously from the evolution operator $S_{\lambda_0}(t)$ of the underlying system their results cover the case where $S_\lambda(t)$ represents certain finite dimensional spatial discretizations of parabolic gradient partial differential equations.

In [36] we consider the numerical solution, by a Runge-Kutta method (3.2.1–2), of gradient systems with hyperbolic equilibria which are dissipative in the sense of (6.1.2), and use the method of proof of Hale and Raugel to show lower semicontinuity at $h = 0$ of the numerical approximation to the global attractor. However this result is *not* reproduced in Section 7.3; the global attractor of a gradient system with hyperbolic equilibria has the form $\mathcal{A} = \overline{W(\mathcal{E}^*)}$ and so this result is just a special case of Theorem 7.3.3.

The method of proof of Hale and Raugel uses the Morse decomposition of an attractor of a gradient system. Our new approach to proving lower semicontinuity does not require that the underlying system is in gradient form, and uses a compactness argument instead of relying on a Morse decomposition. This not only leads to shorter and more elegant proofs but since, for example, the system in Example 2.3.9 is not in gradient form, but nevertheless has a global attractor which is the closure of the (unique) hyperbolic fixed point of the system, we have established lower semicontinuity for numerical approximation to attractors of some non-gradient systems. Moreover,

little is known about the structure of strange attractors, such as the Lorenz attractor, and it is possible that some chaotic systems may have attractors of this form. To the knowledge of the author, this is the first proof of lower semicontinuity for non-gradient systems.

The terms upper and lower semicontinuity come from set-valued analysis. In Section 7.4 we introduce two further concepts from set-valued analysis, namely the $\liminf_{h \rightarrow 0} \mathcal{A}_h$ and $\limsup_{h \rightarrow 0} \mathcal{A}_h$, which enable us to study the behaviour of general numerical invariant sets and attractors \mathcal{A}_h in the limit as $h \rightarrow 0$. We denote the $\liminf_{h \rightarrow 0} \mathcal{A}_h$ and $\limsup_{h \rightarrow 0} \mathcal{A}_h$ by \mathcal{A}_0^- and \mathcal{A}_0^+ respectively. Then \mathcal{A}_0^- and \mathcal{A}_0^+ are by definition closed sets with $\mathcal{A}_0^- \subseteq \mathcal{A}_0^+$, and they are, in some sense, the smallest and largest set of numerically observable invariant dynamics in the limit as $h \rightarrow 0$. Moreover \mathcal{A}_h converges in Hausdorff metric to a compact set \mathcal{A}_0 as $h \rightarrow 0$ if and only if $\mathcal{A}_0^- = \mathcal{A}_0^+ = \mathcal{A}_0$.

We then show that if the \mathcal{A}_h 's are forward or backward invariant then both \mathcal{A}_0^- and \mathcal{A}_0^+ are forward or backward invariant, respectively. Combining these results in Theorem 7.4.7 we prove that if \mathcal{A}_h is invariant under evolution of the numerical approximation for all $h \in (0, h_0)$ and \mathcal{A}_h converges in Hausdorff metric to a compact set \mathcal{A}_0 as $h \rightarrow 0$ then \mathcal{A}_0 is invariant under evolution of the underlying system (7.1.1). This result is more significant than it at first appears. Unlike the other results in this chapter in Theorem 7.4.7 we have made no assumptions on the dynamics of the underlying system (except that \mathbf{f} is locally Lipschitz), and have used the existence of numerically invariant sets to deduce the existence of an invariant set for the underlying system. In contrast many results in the literature assume that the dynamics of the underlying system has a certain form and then prove that this is preserved by the numerical approximation. For example, Beyn [3] and Eirola [14] prove that if (7.1.1) has a hyperbolic periodic orbit then a one-step discretization will also possess an invariant curve, for h sufficiently small, which converges to the periodic orbit of the underlying system as $h \rightarrow 0$. Theorem 7.4.7 can be considered as a converse to such results. Theorem 7.4.7 is also related to results in Chapter 4; it implies an absence of spuriousity in the limit as $h \rightarrow 0$. In Theorem 4.2.8 we showed that a continuous branch of fixed points either becomes unbounded or converges to a fixed point of the underlying system as $h \rightarrow 0$. Theorem 7.4.7 extends this result from applying only to the convergence of fixed points, to apply to the convergence of arbitrary invariant sets whether they be

fixed points, periodic orbits, tori, strange attractors or whatever.

Finally we consider the numerical approximation of a local or global attractor \mathcal{A} once more, but now *without* the assumption that $\mathcal{A} = \overline{W(\mathcal{E}^*)}$, although we still assume that f is \mathcal{C}^1 on a neighbourhood of the hyperbolic fixed points of the \mathcal{A} . In this case $\overline{W(\mathcal{E}^*)} \subset \mathcal{A}$ with strict inclusion, and indeed if \mathcal{A} contains no hyperbolic fixed points then $\overline{W(\mathcal{E}^*)}$ will be empty. However Result 2.6.2 implies that if \mathcal{A} is a global attractor (which has a convex absorbing set) then it will contain at least one fixed point, and since fixed points are generically hyperbolic, $\overline{W(\mathcal{E}^*)}$ will in general be nonempty for a global attractor. In Theorem 7.4.8 we show that \mathcal{A}_0^- and \mathcal{A}_0^+ are both invariant (under the evolution of (7.1.1)) subsets of \mathcal{A}_0 with $\overline{W(\mathcal{E}^*)} \subseteq \mathcal{A}_0^- \subseteq \mathcal{A}_0^+ \subseteq \mathcal{A}$ and hence if the numerical approximations \mathcal{A}_h converge in Hausdorff metric to a compact set \mathcal{A}_0 as $h \rightarrow 0$ then \mathcal{A}_0 is an invariant subset of \mathcal{A} which contains $\overline{W(\mathcal{E}^*)}$.

7.2 Upper Semicontinuity

We begin this section by showing that for h sufficiently small the numerical approximation to (7.1.1) by a Runge-Kutta method (3.2.1–2) defines a discrete dynamical system on a neighbourhood of any attractor \mathcal{A} of the underlying system, and itself possesses a local attractor \mathcal{A}_h in this neighbourhood. We will then go on to show that $\text{dist}(\mathcal{A}_h, \mathcal{A}) \rightarrow 0$ as $h \rightarrow 0$. The proof of Proposition 7.2.1 follows the method of Hale, Lin & Raugel [29] and Temam [58], and basically entails balancing the attraction of the attractor against the numerical errors.

Proposition 7.2.1 *Suppose that (7.1.1) defines a dynamical system on \mathbb{R}^m with local attractor \mathcal{A} which attracts a bounded forward invariant neighbourhood N of itself. If this system is approximated numerically using a Runge-Kutta method (3.2.1–2), where for an implicit method the solution of (3.2.1–2) is as defined by Proposition 3.6.3, then given any $\varepsilon > 0$ there exists $h_0 = h_0(\varepsilon) > 0$ such that for $h \in (0, h_0)$ the numerical solution defines a continuous discrete dynamical system on N_h where $N \subseteq N_h \subseteq \mathcal{N}(N, \varepsilon)$ is defined by (7.2.2), which possesses an attractor $\mathcal{A}_h \subseteq \mathcal{N}(\mathcal{A}, \varepsilon)$ which attracts N_h .*

Proof. Given $\varepsilon > 0$ reduce ε if necessary so that $\mathcal{N}(\mathcal{A}, 2\varepsilon) \subseteq N$. Now since \mathcal{A} attracts N there exists $t^* \geq 0$ such that

$$S(t)N \subseteq \mathcal{N}(\mathcal{A}, \varepsilon/2) \quad \text{for all } t \geq t^*. \quad (7.2.1)$$

Now applying Proposition 3.6.8 with $t_0 = 2t^*$ we deduce the existence of a unique solution sequence $\{\mathbf{y}_n\}_{n=0}^{n^*}$ if $h \in (0, h_0)$ for some $h_0 > 0$, for any $\mathbf{y}_0 \in N$ where n^* is the largest integer such that $n^*h \leq 2t^*$, which satisfies $\mathbf{y}_n \in \mathcal{N}(N, \varepsilon/2)$ and the global error bound (3.6.16). This implies that

$$N_h = \bigcup_{n \leq n^*-1} S_h^n N \quad (7.2.2)$$

is well defined, where S_h the evolution semi-group of the numerical solution is well defined on N_h .

Reducing h_0 if necessary so that

$$h_0 \leq \frac{\varepsilon}{2CN(e^{2Lt^*} - 1)}$$

where L is the Lipschitz constant for \mathbf{f} on $\mathcal{N}(N, \varepsilon)$, equation (3.6.16) implies that the global error satisfies $e_n(h) \leq \varepsilon/2$ for $n \leq n^*$. Since $S(t)\mathbf{y}_0 \in N$ for all $\mathbf{y}_0 \in N$ and all $t \geq 0$ it follows that $N_h \subseteq \mathcal{N}(N, \varepsilon/2)$ as required. Next we show that N_h is forward invariant under S_h .

Suppose that N_h is not forward invariant under S_h , then there exists $\mathbf{x} \in S_h N_h$ such that $\mathbf{x} \notin N_h$. By (7.2.2) such an \mathbf{x} must satisfy $\mathbf{x} \in S_h^{n^*} N$ and hence $\mathbf{x} = S_h^{n^*} \mathbf{y}_0$ for some $\mathbf{y}_0 \in N$. Now since $e_n(h) \leq \varepsilon/2$ for all $\mathbf{y}_0 \in N$ and $n \leq n^*$ it follows that $\|S_h^{n^*} \mathbf{y}_0 - S(n^*h)\mathbf{y}_0\| \leq \varepsilon/2$. But (7.2.1) implies that $S(n^*h)\mathbf{y}_0 \in \mathcal{N}(\mathcal{A}, \varepsilon/2)$ and hence it follows that $\mathbf{x} = S_h^{n^*} \mathbf{y}_0 \in \mathcal{N}(\mathcal{A}, \varepsilon) \subseteq N \subseteq N_h$, which supplies the required contradiction. Thus N_h is forward invariant under S_h , and the numerical solution defines a discrete dynamical system on N_h . Continuity with respect to initial data follows from Theorem 3.6.4.

It remains to show that the numerical solution possesses a local attractor $\mathcal{A}_h \subseteq \mathcal{N}(\mathcal{A}, \varepsilon)$ which attracts N_h . Since N_h is a neighbourhood of $\mathcal{N}(\mathcal{A}, \varepsilon)$, to do this it is sufficient to show that $\mathcal{N}(\mathcal{A}, \varepsilon)$ is an absorbing set for the discrete dynamical system on N_h , which we will now do. Suppose $h \in (0, h_0)$ and integer n satisfies $t_0 \leq nh \leq 2t_0$ then for any $\mathbf{y}_0 \in N_h$

$$\begin{aligned} \text{dist}(S_h^n \mathbf{y}_0, \mathcal{A}) &= \inf_{\mathbf{x} \in \mathcal{A}} \|S_h^n \mathbf{y}_0 - \mathbf{x}\| \\ &\leq \|S_h^n \mathbf{y}_0 - S(nh)\mathbf{y}_0\| + \inf_{\mathbf{x} \in \mathcal{A}} \|S(nh)\mathbf{y}_0 - \mathbf{x}\| \\ &\leq \|S_h^n \mathbf{y}_0 - S(nh)\mathbf{y}_0\| + \text{dist}(S(nh)\mathbf{y}_0, \mathcal{A}) \end{aligned}$$

$$< \varepsilon.$$

Thus $S_h^n N_h \subseteq \mathcal{N}(\mathcal{A}, \varepsilon)$ for all n such that $t_0 \leq nh \leq 2t_0$. We will establish by induction that $S_h^n N_h \subseteq \mathcal{N}(\mathcal{A}, \varepsilon)$ for all integer $n: nh \geq t_0$. Suppose the result holds for integer $n: t_0 \leq nh \leq kt_0$ with $k \geq 2$ and consider integer n such that $kt_0 \leq nh \leq (k+1)t_0$.

Choose m such that $t_0 \leq mh \leq kt_0$ and let $p = n - m$. Then $nh = (m + p)h$ and $0 \leq ph \leq kt_0$. Now $S_h^n N_h = S_h^m S_h^p N_h$ and since N_h is forward invariant it follows that $S_h^p N_h \subseteq N_h$ and hence that $S_h^n N_h \subseteq S_h^m N_h$. Thus since $t_0 \leq mh \leq kt_0$ it follows from the inductive hypothesis that $S_h^n N_h \subseteq S_h^m N_h \subseteq \mathcal{N}(\mathcal{A}, \varepsilon)$ and the induction argument is complete; $S_h^n N_h \subseteq \mathcal{N}(\mathcal{A}, \varepsilon)$ for all integer $n: nh \geq t_0$. Thus $\mathcal{N}(\mathcal{A}, \varepsilon)$ is an absorbing set for the discrete dynamical system defined by the numerical solution on N_h , and thus that this discrete dynamical system possesses a local attractor \mathcal{A}_h with $\mathcal{A}_h \subseteq \mathcal{N}(\mathcal{A}, \varepsilon)$ as required. ■

Upper semicontinuity at $h = 0$ for numerical approximation of both local and global attractors follows trivially from Proposition 7.2.1 on noting that $\mathcal{A}_h \subseteq \mathcal{N}(\mathcal{A}, \varepsilon)$ implies that $\text{dist}(\mathcal{A}_h, \mathcal{A}) \leq \varepsilon$.

Theorem 7.2.2 *Suppose that (7.1.1) defines a dynamical system on \mathbb{R}^m with an attractor \mathcal{A} which attracts a bounded forward invariant neighbourhood N of itself. If this system is approximated numerically using a Runge-Kutta method (3.2.1–2), where for an implicit method the solution of (3.2.1–2) is as defined by Proposition 3.6.3, then there exists $h_0 > 0$ such that for $h \in (0, h_0)$ the numerical solution defines a continuous discrete dynamical system on $N_h \supseteq N$, which possesses an attractor \mathcal{A}_h that attracts N_h (and hence N), and satisfies*

$$\text{dist}(\mathcal{A}_h, \mathcal{A}) \rightarrow 0 \quad \text{as } h \rightarrow 0. \quad \blacksquare$$

Remark (i) \mathcal{A} may be a local or a global attractor. In the case of a global attractor it follows from Theorem 7.2.2 that N may be arbitrarily large.

(ii) When approximating a dissipative system with a global attractor \mathcal{A} Theorem 7.2.2 only ensures that the numerical approximation possesses a local attractor \mathcal{A}_h . In general \mathcal{A}_h will not be a global attractor, however the domain of attraction of \mathcal{A}_h can be made arbitrarily large by taking h sufficiently small. Moreover, by imposing more structure on the problem, such as (6.1.2), and solving with an algebraically stable

method, as in Chapter 6, we can ensure that \mathcal{A}_h is globally attracting.

(iii) Theorem 7.2.2 implies that numerical attractors do not contain any spurious features in the limit as $h \rightarrow 0$, since given any $\varepsilon > 0$ there exists $h_0 = h_0(\varepsilon) > 0$ such that for $h \in (0, h_0)$ $\mathcal{A}_h \subset \mathcal{N}(\mathcal{A}, \varepsilon)$.

(iii) However, these results do not imply that the numerical attractor \mathcal{A}_h reproduces all the features of the underlying attractor \mathcal{A} . We have not considered $\text{dist}(\mathcal{A}, \mathcal{A}_h)$, and this semi-distance might be large. We will consider the behaviour of $\text{dist}(\mathcal{A}, \mathcal{A}_h)$ as $h \rightarrow 0$ in the remaining sections.

7.3 Lower Semicontinuity

Recall from Theorem 2.3.8 that for a dissipative system $\overline{W(\mathcal{E}^*)}$, the closure of the union of the unstable manifolds of the hyperbolic fixed points of the system, is a compact, invariant subset of the global attractor \mathcal{A} . In this section we will consider both global and local attractors under the assumption that they have the form $\mathcal{A} = \overline{W(\mathcal{E}^*)}$ and prove that under numerical approximation by a Runge-Kutta method the corresponding discrete attractor \mathcal{A}_h (the existence of which, for h sufficiently small, was proved in the last section) converges to the attractor \mathcal{A} of the underlying system in the Hausdorff metric as $h \rightarrow 0$.

Recall from Section 2.3 that the global attractor of a dissipative gradient system with hyperbolic equilibria has the form $\mathcal{A} = \overline{W(\mathcal{E}^*)}$, and so our result applies to all such systems. Note also that the global attractor of the system in Example 2.3.9 has the same form, although this system is not in gradient form. Our result then, applies to the attractors of at least some nongradient systems. The known method of proof of lower semicontinuity due to Hale and Raugel [30] applies only to gradient systems, and so the new method of proof given below, as well as being more elegant than that of Hale and Raugel, actually applies to a wider class of systems. We will consider the more general case where $\overline{W(\mathcal{E}^*)} \subset \mathcal{A}$, with strict inclusion, at the end of Section 7.4.

Since we proved that $\text{dist}(\mathcal{A}_h, \mathcal{A}) \rightarrow 0$ as $h \rightarrow 0$ in the last section, to prove convergence of \mathcal{A}_h to \mathcal{A} in the Hausdorff metric it only remains to show that

$$\text{dist}(\mathcal{A}, \mathcal{A}_h) \rightarrow 0 \quad \text{as } h \rightarrow 0.$$

Our approach to proving this when $\mathcal{A} = \overline{W(\mathcal{E}^*)}$ is simply to show that the unstable

manifolds of the hyperbolic fixed points of the underlying system are well approximated by their numerical counterparts. The proof of this relies on a result of Beyn [4], who proves such a result for local unstable manifolds of hyperbolic fixed points.

Beyn [4] considers the numerical approximation of dynamical systems in a neighbourhood of a hyperbolic fixed point by general one-step and multistep methods. He showed that the phase portraits of the underlying system are correctly reproduced by the numerical approximation, on a sufficiently small neighbourhood of a hyperbolic fixed point. Amongst other implications, this implies that the local stable and unstable manifolds of the fixed point are well approximated numerically.

The following result is a special case of Theorem 3.1 of Beyn [4]. It shows that we can approximate the local unstable manifold of a hyperbolic fixed to arbitrary precision (in the Hausdorff metric) when discretizing with a Runge-Kutta method. The result stated in Beyn [4] is actually far more general than stated here, but we do not require Beyn's theorem in its full generality, and so only state the special case that we need. Recall that unstable and local unstable manifolds were formally defined in Definition 2.3.7.

Result 7.3.1 (Beyn [4]) *If $\mathbf{x}^* \in U$ is a hyperbolic equilibrium of (7.1.1), where $\mathbf{f} \in C^1(U, \mathbb{R}^m)$ and (7.1.1) defines a dynamical system on U then for any Runge-Kutta method (3.2.1–2), where the solution of (3.2.1) is as defined by Proposition 3.6.3, there exists $\Delta > 0$ such that for $0 < \delta < \Delta$ and any $\varepsilon > 0$ there exists $h_0 = h_0(\delta, \varepsilon) > 0$ such that if $h \in (0, h_0)$ then*

$$\text{dist}_H(W^\delta(\mathbf{x}^*), W_h^\delta(\mathbf{x}^*)) < \varepsilon$$

and in particular

$$\text{dist}(W^\delta(\mathbf{x}^*), W_h^\delta(\mathbf{x}^*)) < \varepsilon. \quad \blacksquare \quad (7.3.1)$$

We now show that the closure of the union of the unstable manifolds of the hyperbolic equilibria of an attractor, $\overline{W(\mathcal{E}^*)}$, is well approximated by its numerical counterpart $\overline{W_h(\mathcal{E}^*)}$. The proof of this result essentially consists of applying a compactness argument to Beyn's similar result for the local unstable manifolds, although the technicalities in the proof make it appear more complicated than it really is.

Proposition 7.3.2 *Suppose that (7.1.1) defines a dynamical system on \mathbb{R}^m with local attractor \mathcal{A} which attracts a bounded open forward invariant neighbourhood N of itself. Let \mathcal{E}^* be the set of hyperbolic fixed points contained in the attractor and $W(\mathcal{E}^*)$ the*

union of the unstable manifolds of these hyperbolic fixed points. If f is C^1 on a neighbourhood of \mathcal{E}^* then the continuous discrete dynamical system on N_h defined, as in Proposition 7.2.1, by approximating (7.1.1) numerically using a Runge-Kutta method (3.2.1–2) satisfies

$$\text{dist}(\overline{W(\mathcal{E}^*)}, \overline{W_h(\mathcal{E}^*)}) \rightarrow 0 \quad \text{as } h \rightarrow 0 \quad (7.3.2)$$

where $W_h(\mathcal{E}^*)$ is the numerical counterpart to $W(\mathcal{E}^*)$.

Proof. It is sufficient to show that given any $\varepsilon > 0$ there exists $h_0 = h_0(\varepsilon) > 0$ such that if $h \in (0, h_0)$ then $\text{dist}(\overline{W(\mathcal{E}^*)}, \overline{W_h(\mathcal{E}^*)}) \leq \varepsilon$. But note that if $\overline{W(\mathcal{E}^*)} \subseteq \mathcal{N}(\overline{W_h(\mathcal{E}^*)}, \varepsilon)$ then $\text{dist}(\overline{W(\mathcal{E}^*)}, \overline{W_h(\mathcal{E}^*)}) \leq \varepsilon$, hence we merely need to show that given any $\varepsilon > 0$ there exists $h_0 = h_0(\varepsilon) > 0$ such that if $h \in (0, h_0)$ then $\overline{W(\mathcal{E}^*)} \subseteq \mathcal{N}(\overline{W_h(\mathcal{E}^*)}, \varepsilon)$.

Consider an open cover

$$\{B(\mathbf{x}, \varepsilon/4) : \mathbf{x} \in W(\mathcal{E}^*)\}$$

of $W(\mathcal{E}^*)$. Clearly this also covers $\overline{W(\mathcal{E}^*)}$ and so since $\overline{W(\mathcal{E}^*)}$ is compact there exists a finite subcover

$$\{B(\mathbf{x}_i, \varepsilon/4) : i = 1, \dots, n, \mathbf{x}_i \in W(\mathcal{E}^*)\}$$

of $\overline{W(\mathcal{E}^*)}$. Let $C = \{\mathbf{x}_i : i = 1, \dots, n\}$, the set of centres of the covering balls.

Claim: There exists $h_0 = h_0(\varepsilon) > 0$ such that if $h \in (0, h_0)$ then for each $\mathbf{x}_i \in C$ there exists $\mathbf{y}_i \in W_h(\mathcal{E}^*)$ such that $\|\mathbf{x}_i - \mathbf{y}_i\| \leq \frac{\varepsilon}{2}$.

The result follows easily from this claim, as we now show. Suppose $\mathbf{x} \in \overline{W(\mathcal{E}^*)}$ then

$$\text{dist}(\mathbf{x}, \overline{W_h(\mathcal{E}^*)}) \leq \text{dist}(\mathbf{x}, \mathbf{x}_i) + \text{dist}(\mathbf{x}_i, \overline{W_h(\mathcal{E}^*)})$$

for any $\mathbf{x}_i \in C$, by the triangle inequality. But since $\{B(\mathbf{x}_i, \varepsilon/4) : \mathbf{x}_i \in C\}$ covers $\overline{W(\mathcal{E}^*)}$ it follows that $\mathbf{x} \in B(\mathbf{x}_i, \varepsilon/4)$ for some i , and that $\text{dist}(\mathbf{x}, \mathbf{x}_i) \leq \varepsilon/4$ for this i . Now using the claim

$$\begin{aligned} \text{dist}(\mathbf{x}_i, \overline{W_h(\mathcal{E}^*)}) &\leq \text{dist}(\mathbf{x}_i, \mathbf{y}_i) \\ &\leq \frac{\varepsilon}{2}. \end{aligned}$$

Therefore $\text{dist}(\mathbf{x}, \overline{W_h(\mathcal{E}^*)}) < \varepsilon$ and hence $\mathbf{x} \in \mathcal{N}(\overline{W_h(\mathcal{E}^*)}, \varepsilon)$. But this is true for all $\mathbf{x} \in \overline{W(\mathcal{E}^*)}$, and hence $\overline{W(\mathcal{E}^*)} \subseteq \mathcal{N}(\overline{W_h(\mathcal{E}^*)}, \varepsilon)$ as required. This completes the proof,

subject to the claim.

Proof of Claim: Consider \mathcal{E}^* the set of hyperbolic fixed points. Since hyperbolicity of a fixed point implies that it is isolated and \mathcal{E}^* is bounded it follows that \mathcal{E}^* is a finite set;

$$\mathcal{E}^* = \{\mathbf{x}_j^* : j = 1, \dots, m\}.$$

Since \mathbf{f} is \mathcal{C}^1 on a neighbourhood of \mathcal{E}^* , for each $\mathbf{x}_j^* \in \mathcal{E}^*$ there exists $\Delta_j > 0$ such that the conclusions of Result 7.3.1 hold on $B(\mathbf{x}_j^*, \Delta_j)$. Let $\delta = \frac{1}{2} \min_j \Delta_j$.

Choose arbitrary $h_0 > 0$.

Since $\mathbf{x}_i \in C$ satisfies $\mathbf{x}_i \in \overline{W(\mathcal{E}^*)}$ there exists $\mathbf{x}^* \in \mathcal{E}^*$ such that $\mathbf{x}_i \in W(\mathbf{x}^*)$. It follows from the definition of local unstable manifolds that there exists $\bar{\mathbf{x}}_i \in W^\delta(\mathbf{x}^*)$ and (finite) $t^* \geq 0$ such that $\mathbf{x}_i = S(t^*)\bar{\mathbf{x}}_i$.

Applying Result 7.3.1 we deduce the existence of $h_i > 0$ such that for $h \in (0, h_i)$

$$\text{dist}(W^\delta(\mathbf{x}^*), W_h^\delta(\mathbf{x}^*)) \leq \frac{\varepsilon}{4e^{L(t^*+h_0)}} \quad (7.3.3)$$

where L is the Lipschitz constant for \mathbf{f} on $\mathcal{N}(N, \varepsilon)$. If $h_0 > h_i$ then reduce h_0 by setting $h_0 = h_i$.

Note that given any $h \in (0, h_0)$ there exists of $\tilde{\mathbf{x}}_i \in W^\delta(\mathbf{x}^*)$ such that $\mathbf{x}_i = S(kh)\tilde{\mathbf{x}}_i$ where k is a positive integer such that $kh \in [t^*, t^* + h_0]$.

Now by (7.3.3) there exists $\tilde{\mathbf{y}}_i \in W_h^\delta(\mathbf{x}^*)$ such that

$$\|\tilde{\mathbf{x}}_i - \tilde{\mathbf{y}}_i\| \leq \frac{\varepsilon}{4e^{L(t^*+h_0)}}.$$

Hence by Result 2.1.4

$$\begin{aligned} \|S(kh)\tilde{\mathbf{y}}_i - \mathbf{x}_i\| &= \|S(kh)\tilde{\mathbf{y}}_i - S(kh)\tilde{\mathbf{x}}_i\| \\ &\leq \frac{\varepsilon}{4}. \end{aligned} \quad (7.3.4)$$

Now Proposition 3.6.8 implies that there exists $h_i > 0$ such that if $h \in (0, h_i)$ then $S_h^k \tilde{\mathbf{y}}_i$ is well defined and

$$\|S_h^k \tilde{\mathbf{y}}_i - S(kh)\tilde{\mathbf{y}}_i\| \leq \frac{\varepsilon}{4}. \quad (7.3.5)$$

Again, if $h_0 > h_i$ then reduce h_0 by setting $h_0 = h_i$.

Now let $\mathbf{y}_i = S_h^k \tilde{\mathbf{y}}_i$, and note that this implies that $\mathbf{y}_i \in W_h(\mathcal{E}^*)$. Moreover using

method the solution of (3.2.1–2) is as defined by Proposition 3.6.3, then there exists $h_0 > 0$ such that for $h \in (0, h_0)$ the numerical solution defines a continuous discrete dynamical system on $N_h \supseteq N$, which possesses an attractor \mathcal{A}_h that attracts N_h (and hence N), and satisfies

$$\text{dist}_H(\mathcal{A}_h, \mathcal{A}) \rightarrow 0 \quad \text{as } h \rightarrow 0.$$

Proof. Follows trivially from Theorem 7.2.2 together with Proposition 7.3.2. ■

Remarks (i) \mathcal{A} may be a local or a global attractor. In the case of a global attractor it follows from Theorem 7.2.2 that N may be arbitrarily large.

(ii) Note that $\mathcal{A} = \overline{W(\mathcal{E}^*)}$ does not imply that $\mathcal{A}_h = \overline{W_h(\mathcal{E}^*)}$. We have not considered the form of the discrete attractor \mathcal{A}_h ; it is irrelevant in our method of proof.

(iii) For the theorem to be of much practical use we need to determine when the attractor \mathcal{A} has the required form. However, for general chaotic dynamical systems this question is not easy to answer.

(iv) The method of proof of lower semicontinuity of Hale and Raugel only applies to gradient dynamical systems with hyperbolic equilibria. For such systems $\mathcal{A} = \overline{W(\mathcal{E}^*)}$ and Theorem 7.3.3 applies. Since the attractor in Example 2.3.9 also has this form, although the system is not in gradient form, our new method of proof does generalize the class of systems to which the result applies, even if it is hard to identify which dynamical systems have attractors of this form.

(v) There may be chaotic attractors which are of the form $\mathcal{A} = \overline{W(\mathcal{E}^*)}$. Since unstable periodic orbits are typically dense in a chaotic attractor and $W(\mathcal{E}^*)$ cannot contain periodic orbits this may at first seem to be untrue, but Example 2.3.9 shows that $\overline{W(\mathcal{E}^*)}$ can contain periodic orbits.

(vi) Hyperbolic periodic orbits also persist under numerical approximation [3, 14]. If we let \mathcal{P}^* be the set of points on the attractor that are on hyperbolic periodic orbits then the theorem can easily be extended to attractors that are of the form $\mathcal{A} = \overline{W(\mathcal{E}^*) \cup \mathcal{P}^*}$. If a result similar to Result 7.3.1 for the numerical approximation to the local unstable manifolds of hyperbolic periodic orbits then Theorem 7.3.3 could be further extended to cover attractors of the form $\mathcal{A} = \overline{W(\mathcal{E}^*) \cup W(\mathcal{P}^*)}$.

7.4 Invariant Sets and Attractors

In this section we will consider the convergence of numerical invariant sets \mathcal{A}_h as $h \rightarrow 0$. We will begin by introducing concepts of \liminf and \limsup for a continuum of sets and go on to prove two main results. In Theorem 7.4.7 we show that if $\{\mathcal{A}_h\}_{h \in (0, h_0)}$ is uniformly bounded, each \mathcal{A}_h is invariant under evolution of the discretized system, and \mathcal{A}_h converges in Hausdorff metric to a compact set \mathcal{A}_0 as $h \rightarrow 0$ then \mathcal{A}_0 is an invariant set of the underlying system. In Theorem 7.4.8 we show that if the underlying system possesses an attractor \mathcal{A} and its numerical approximations \mathcal{A}_h converge in Hausdorff metric to a compact set \mathcal{A}_0 as $h \rightarrow 0$ then \mathcal{A}_0 is an invariant subset of the \mathcal{A} which contains $\overline{W(\mathcal{E}^*)}$. We also consider the cases where \mathcal{A}_h does not converge in Hausdorff metric as $h \rightarrow 0$ in the setting of both theorems.

Just as a sequence in \mathbb{R} need not converge to a limit in \mathbb{R} there is no reason why a sequence or continuum of sets $\{\mathcal{A}_h\}_{h \in (0, h_0)}$ need converge (in any set metric) as $h \rightarrow 0$. For this reason, we need concepts of $\limsup_{h \rightarrow 0} \mathcal{A}_h$ and $\liminf_{h \rightarrow 0} \mathcal{A}_h$ with analogous properties to \liminf 's and \limsup 's of sequences in \mathbb{R} . We make similar definitions to those on page 41 of [2].

Definition 7.4.1 Given $\{\mathcal{A}_h\}_{h \in (0, h_0)}$ where each $\mathcal{A}_h \subset \mathbb{R}^m$ we define the sets \mathcal{A}_0^+ and \mathcal{A}_0^- by

$$\begin{aligned} \mathcal{A}_0^+ &= \limsup_{h \rightarrow 0} \mathcal{A}_h. \\ &:= \left\{ \mathbf{x} : \liminf_{h \rightarrow 0} \text{dist}(\mathbf{x}, \mathcal{A}_h) = 0 \right\} \end{aligned} \quad (7.4.1)$$

$$= \left\{ \mathbf{x} : \begin{array}{l} \exists \{h_i\}_{i=1}^\infty \ \& \ \{\mathbf{x}_i\}_{i=1}^\infty \text{ such that } h_i \in (0, h_0), \mathbf{x}_i \in \mathcal{A}_{h_i} \\ \mathbf{x}_i \rightarrow \mathbf{x} \ \& \ h_i \rightarrow 0 \text{ (monotonically) as } i \rightarrow \infty \end{array} \right\} \quad (7.4.2)$$

and

$$\begin{aligned} \mathcal{A}_0^- &= \liminf_{h \rightarrow 0} \mathcal{A}_h \\ &:= \left\{ \mathbf{x} : \lim_{h \rightarrow 0} \text{dist}(\mathbf{x}, \mathcal{A}_h) = 0 \right\} \end{aligned} \quad (7.4.3)$$

$$= \left\{ \mathbf{x} : \begin{array}{l} \forall \{h_i\}_{i=1}^\infty \text{ satisfying } h_i \in (0, h_0), h_i \rightarrow 0 \text{ (monotonically),} \\ \text{there exists } \{\mathbf{x}_i\}_{i=1}^\infty \text{ such that } \mathbf{x}_i \in \mathcal{A}_{h_i} \ \& \ \mathbf{x}_i \rightarrow \mathbf{x} \end{array} \right\} \quad (7.4.4)$$

Clearly $\mathcal{A}_0^- \subseteq \mathcal{A}_0^+$, and both \mathcal{A}_0^- and \mathcal{A}_0^+ are closed.

It is impossible to perform uncountably many numerical simulations to produce a continuum of numerical approximations $\{\mathcal{A}_h\}_{h \in (0, h_0)}$ to an invariant set \mathcal{A} of some underlying simulation. However we can (at least visualize) taking a countable sequence of numerical approximations \mathcal{A}_{h_i} with $h_i \rightarrow 0$ as $i \rightarrow \infty$.

Then \mathcal{A}_0^+ , as defined by (7.4.2), is the union over all monotonic sequences $h_i \rightarrow 0$ of the points which can be reached as a limit of points in \mathcal{A}_{h_i} as $i \rightarrow \infty$. As such \mathcal{A}_0^+ contains all the possible ‘invariant dynamics’ which *might be* observed in the limit as $h_i \rightarrow 0$ when $\{\mathcal{A}_h\}_{h \in (0, h_0)}$ is sampled over a sequence $h_i \rightarrow 0$.

In contrast \mathcal{A}_0^- , as defined by (7.4.4), is the intersection over all monotonic sequences $h_i \rightarrow 0$ of the points which can be reached as a limit of points in \mathcal{A}_{h_i} as $i \rightarrow \infty$. As such \mathcal{A}_0^- contains all the ‘invariant dynamics’ which *must be* observed in the limit as $h_i \rightarrow 0$ when $\{\mathcal{A}_h\}_{h \in (0, h_0)}$ is sampled over any sequence $h_i \rightarrow 0$.

We still need to give more than a little justification for calling \mathcal{A}_0^+ and \mathcal{A}_0^- a limsup and a liminf respectively. We will do this by showing that \mathcal{A}_h converges to a compact set \mathcal{A} in the Hausdorff metric as $h \rightarrow 0$ if and only if $\mathcal{A}_0^- = \mathcal{A}_0^+ = \mathcal{A}$. We now state two propositions which will enable us to do this.

Proposition 7.4.2 *Given $\{\mathcal{A}_h\}_{h \in (0, h_0)}$ where the \mathcal{A}_h ’s are uniformly bounded in \mathbb{R}^m (i.e. there exists bounded $B \subset \mathbb{R}^m$ such that $\mathcal{A}_h \subset B$ for all $h \in (0, h_0)$) and \mathcal{A}_0^+ as defined in Definition 7.4.1 it follows that*

$$\text{dist}(\mathcal{A}_h, \mathcal{A}_0^+) \rightarrow 0 \quad \text{as} \quad h \rightarrow 0. \quad (7.4.5)$$

This implies that the following statements are equivalent.

- (i) $\text{dist}(\mathcal{A}_h, \mathcal{A}) \rightarrow 0$ as $h \rightarrow 0$,
- (ii) $\mathcal{A}_0^+ \subseteq \overline{\mathcal{A}}$.

Proof. To show that (7.4.5) holds it is sufficient to show that given any $\varepsilon > 0$ there exists $h_1 > 0$ such that if $h \in (0, h_1)$ then $\mathcal{A}_h \subseteq \mathcal{N}(\mathcal{A}_0^+, \varepsilon)$. Suppose this fails for some $\varepsilon > 0$. Then for this ε we can construct sequences $\{h_i\}_{i=1}^\infty$ and $\{\mathbf{x}_i\}_{i=1}^\infty$ such that $h_i \rightarrow 0$ as $i \rightarrow \infty$ and $\mathbf{x}_i \in \mathcal{A}_{h_i}$ and $\mathbf{x}_i \notin \mathcal{N}(\mathcal{A}_0^+, \varepsilon)$ for all i . But since the \mathcal{A}_{h_i} are uniformly bounded we can choose a convergent subsequence of the \mathbf{x}_i ’s. By definition the point to which this subsequence converges is contained in \mathcal{A}_0^+ , but by construction it is not contained in $\mathcal{N}(\mathcal{A}_0^+, \varepsilon)$ which supplies the required contradiction, and thus

(7.4.5) holds.

To show that (i) implies (ii). Suppose (ii) does not hold, then there exists $\mathbf{x} \in \mathcal{A}_0^+$ such that $\mathbf{x} \notin \overline{\mathcal{A}}$. Since $\overline{\mathcal{A}}$ is closed, $\text{dist}(\mathbf{x}, \overline{\mathcal{A}}) > 0$. But now the existence of sequences $\{h_i\}_{i=1}^\infty$ and $\{\mathbf{x}_i\}_{i=1}^\infty$ such that $\mathbf{x}_i \rightarrow \mathbf{x}$, $h_i \rightarrow 0$ and $\mathbf{x}_i \in \mathcal{A}_{h_i}$ for all i contradicts (i).

To see that (ii) implies (i) simply note $\mathcal{A}_0^+ \subseteq \overline{\mathcal{A}}$ implies that

$$\text{dist}(\mathcal{A}_h, \mathcal{A}) = \text{dist}(\mathcal{A}_h, \overline{\mathcal{A}}) \leq \text{dist}(\mathcal{A}_h, \mathcal{A}_0^+)$$

and hence the result follows from (7.4.5). ■

Proposition 7.4.3 *Given $\{\mathcal{A}_h\}_{h \in (0, h_0)}$ where the \mathcal{A}_h 's are uniformly bounded in \mathbb{R}^m (i.e. there exists bounded $B \subset \mathbb{R}^m$ such that $\mathcal{A}_h \subset B$ for all $h \in (0, h_0)$) and \mathcal{A}_0^- as defined in Definition 7.4.1 it follows that*

$$\text{dist}(\mathcal{A}_0^-, \mathcal{A}_h) \rightarrow 0 \quad \text{as} \quad h \rightarrow 0. \quad (7.4.6)$$

This implies that the following statements are equivalent.

- (i) $\text{dist}(\mathcal{A}, \mathcal{A}_h) \rightarrow 0$ as $h \rightarrow 0$,
- (ii) $\overline{\mathcal{A}} \subseteq \mathcal{A}_0^-$.

Proof. To show that (7.4.6) holds it is sufficient to show that given any $\varepsilon > 0$ there exists $h_1 > 0$ such that if $h \in (0, h_1)$ then $\mathcal{A}_0^- \subset \mathcal{N}(\mathcal{A}_h, \varepsilon)$. Suppose this fails for some $\varepsilon > 0$. Then for this ε we can construct sequences $\{h_i\}_{i=1}^\infty$ and $\{\mathbf{x}_i\}_{i=1}^\infty$ such that $h_i \rightarrow 0$ as $i \rightarrow \infty$ and $\mathbf{x}_i \in \mathcal{A}_0^-$ and $\mathbf{x}_i \notin \mathcal{N}(\mathcal{A}_{h_i}, \varepsilon)$ for all i . But since the \mathcal{A}_{h_i} are uniformly bounded it follows trivially that \mathcal{A}_0^- is bounded and we can choose a convergent subsequence of the \mathbf{x}_i 's converging to $\mathbf{x} \in \mathcal{A}_0^-$, say. But now $\mathbf{x}_i \notin \mathcal{N}(\mathcal{A}_{h_i}, \varepsilon)$ for all i implies that for this sequence $\{h_i\}_{i=1}^\infty$ we cannot construct a sequence $\{\mathbf{y}_i\}_{i=1}^\infty$ such that $\mathbf{y}_i \in \mathcal{A}_{h_i}$ and $\mathbf{y}_i \rightarrow \mathbf{x}$ as $i \rightarrow \infty$, which contradicts the fact that $\mathbf{x} \in \mathcal{A}_0^-$. Thus (7.4.6) holds.

To show that (i) implies (ii). Note that (i) implies that $\text{dist}(\overline{\mathcal{A}}, \mathcal{A}_h) \rightarrow 0$ as $h \rightarrow 0$. Now suppose that $\mathbf{x} \in \overline{\mathcal{A}}$. Then for any $\varepsilon > 0$ there exists $h_1 = h_1(\varepsilon) > 0$ such that if $h \in (0, h_1)$ then $B(\mathbf{x}, \varepsilon) \cap \mathcal{A}_h \neq \emptyset$. Thus given any sequence $\{h_i\}_{i=1}^\infty$ such that $h_i \rightarrow 0$ (monotonically) we can construct a sequence $\{\mathbf{x}_i\}_{i=1}^\infty$ such that $\mathbf{x}_i \in \mathcal{A}_{h_i}$ and $\mathbf{x}_i \rightarrow \mathbf{x}$ as $i \rightarrow \infty$. Thus $\mathbf{x} \in \mathcal{A}_0^-$ and (i) implies (ii).

To see that (ii) implies (i) simply note $\overline{\mathcal{A}} \subseteq \mathcal{A}_0^-$ implies that

$$\text{dist}(\mathcal{A}, \mathcal{A}_h) = \text{dist}(\overline{\mathcal{A}}, \mathcal{A}_h) \leq \text{dist}(\mathcal{A}_0^-, \mathcal{A}_h)$$

and hence the result follows from (7.4.6). ■

Combining Propositions 7.4.2 and 7.4.3 we obtain the following theorem.

Theorem 7.4.4 *Given $\{\mathcal{A}_h\}_{h \in (0, h_0)}$ where the \mathcal{A}_h 's are uniformly bounded in \mathbb{R}^m , then with \mathcal{A}_0^- and \mathcal{A}_0^+ as defined in Definition 7.4.1 the following are equivalent.*

- (i) $\text{dist}_H(\mathcal{A}, \mathcal{A}_h) \rightarrow 0$ as $h \rightarrow 0$,
- (ii) $\mathcal{A}_0^- = \mathcal{A}_0^+ = \overline{\mathcal{A}}$.

Proof. To show that (i) implies (ii). Note that $\text{dist}_H(\mathcal{A}, \mathcal{A}_h) \rightarrow 0$ implies that $\text{dist}_H(\overline{\mathcal{A}}, \mathcal{A}_h) \rightarrow 0$. Then it follows from Propositions 7.4.2 and 7.4.3 that $\overline{\mathcal{A}} \subseteq \mathcal{A}_0^- \subseteq \mathcal{A}_0^+ \subseteq \overline{\mathcal{A}}$, and (ii) follows.

To show that (ii) implies (i). Note that, from Proposition 7.4.2, $\mathcal{A}_0^+ = \overline{\mathcal{A}}$ implies that $\text{dist}(\mathcal{A}_h, \mathcal{A}) \rightarrow 0$ as $h \rightarrow 0$, and, from Proposition 7.4.3, $\mathcal{A}_0^- = \overline{\mathcal{A}}$ implies that $\text{dist}(\mathcal{A}, \mathcal{A}_h) \rightarrow 0$ as $h \rightarrow 0$. Thus (i) follows. ■

Remark This result is essentially equivalent to Corollary 1 on page 67 of [1], although the context of the two results is very different.

Thus we have $\mathcal{A}_0^- = \mathcal{A}_0^+$ if and only if \mathcal{A}_h converges to $\mathcal{A}_0 = \mathcal{A}_0^+ = \mathcal{A}_0^-$ in the Hausdorff metric. Moreover $\mathcal{A}_0^- \subseteq \mathcal{A}_0^+$ always holds, and if thus if the inclusion is strict then \mathcal{A}_h does not converge in the Hausdorff metric as $h \rightarrow 0$. This justifies the description of \mathcal{A}_0^- and \mathcal{A}_0^+ as a liminf and a limsup of a continuum of sets respectively; they behave analogously to liminf's and limsup's for sequences in \mathbb{R} (with the partial ordering " \subseteq " for sets replacing the ordering " \leq " for \mathbb{R}).

In Chapters 5 and 6 and in the previous sections of this chapter we have always assumed that the underlying dynamical system has a certain structure, and then shown under what conditions the numerical solution preserved this structure. Whilst such results are important, they assume some knowledge of the behaviour or structure of the underlying system. If possible it is highly desirable to prove results where no assumptions are made on the dynamics or structure of the underlying system but where

the dynamics of the discretized solutions are used to infer properties of the underlying system, and we will now prove some results along these lines.

So far in this section we have considered $\{\mathcal{A}_h\}_{h \in (0, h_0)}$ to be an arbitrary collection of sets. We now suppose that \mathcal{A}_h 's are generated by numerical approximation to (7.1.1) by a Runge-Kutta method with step-size h . The following proposition shows that if \mathcal{A}_h is forward invariant in the discretized system for all $h \in (0, h_0)$ then \mathcal{A}_0^- and \mathcal{A}_0^+ are both forward invariant under the evolution of the underlying system (7.1.1).

Proposition 7.4.5 *Suppose that (7.1.1) defines a dynamical system on \mathbb{R}^m with f locally Lipschitz and that this system is approximated numerically using a Runge-Kutta method (3.2.1–2), where for an implicit method the solution of (3.2.1–2) is as defined by Proposition 3.6.3. Then given a continuum of sets $\{\mathcal{A}_h\}_{h \in (0, h_0)}$ where the \mathcal{A}_h 's are uniformly bounded in \mathbb{R}^m and each \mathcal{A}_h is forward invariant under S_h the evolution operator for the discretized system, then with \mathcal{A}_0^- and \mathcal{A}_0^+ as defined in Definition 7.4.1 it follows that \mathcal{A}_0^+ and \mathcal{A}_0^- are both forward invariant under $S(\bullet)$, the evolution operator for the underlying system.*

Proof. First we show that \mathcal{A}_0^+ is forward invariant. The uniform boundedness of the \mathcal{A}_h 's implies that \mathcal{A}_0^+ is bounded, and that we can choose $\delta > 0$ such that $\mathcal{A}_h \subseteq \mathcal{N}(\mathcal{A}_0^+, \frac{1}{2}\delta)$ for all $h \in (0, h_0)$. Now since f is locally Lipschitz, $S(t)\mathbf{x}$ can only cease to exist if it becomes unbounded. Thus if \mathcal{A}_0^+ is not forward invariant we can choose $\mathbf{x}^* \in \mathcal{A}_0^+$ and $t^* > 0$ such that for all $\mathbf{x} \in \mathcal{N}(\mathcal{A}_0^+, \frac{1}{2}\delta)$ and all $t \in [0, t^*]$ $S(t)\mathbf{x}$ is well defined, $S(t)\mathbf{x} \in \mathcal{N}(\mathcal{A}_0^+, \delta)$ and $S(t^*)\mathbf{x}^* \notin \mathcal{A}_0^+$. With such a choice of \mathbf{x}^* and t^* we will derive a contradiction.

Since $\mathbf{x}^* \in \mathcal{A}_0^+$ there exists a monotonic decreasing sequence $\{h_i\}_{i=1}^\infty$ with $h_i \rightarrow 0$ as $i \rightarrow \infty$ and $\{\mathbf{x}_i\}_{i=1}^\infty$ with $\mathbf{x}_i \in \mathcal{A}_{h_i}$ and $\mathbf{x}_i \rightarrow \mathbf{x}^*$. Choose k_i to be the unique integer such that $k_i h_i \in (t^* - h_i, t^*]$. We now show that

$$\|S_{h_i}^{k_i} \mathbf{x}_i - S(t^*) \mathbf{x}^*\| \rightarrow 0 \quad \text{as} \quad i \rightarrow \infty. \quad (7.4.7)$$

Since each \mathcal{A}_{h_i} is forward invariant $S_{h_i}^{k_i} \mathbf{x}_i \in \mathcal{A}_{h_i}$ and then (7.4.7) implies that $S(t^*) \mathbf{x}^* \in \mathcal{A}_0^+$, the contradiction that we require.

To show that (7.4.7) holds, note that

$$\begin{aligned} \|S_{h_i}^{k_i} \mathbf{x}_i - S(t^*) \mathbf{x}^*\| &\leq \|S_{h_i}^{k_i} \mathbf{x}_i - S(k_i h_i) \mathbf{x}_i\| + \|S(k_i h_i) \mathbf{x}_i - S(k_i h_i) \mathbf{x}^*\| \\ &\quad + \|S(k_i h_i) \mathbf{x}^* - S(t^*) \mathbf{x}^*\|. \end{aligned}$$

By construction $(t^* - k_i h_i) \rightarrow 0$ as $i \rightarrow \infty$ and so $\|S(k_i h_i) \mathbf{x}^* - S(t^*) \mathbf{x}^*\| \rightarrow 0$ as $i \rightarrow \infty$. Since $\mathbf{x}_i \rightarrow \mathbf{x}$ by continuity with respect to initial data $\|S(k_i h_i) \mathbf{x}_i - S(k_i h_i) \mathbf{x}^*\| \rightarrow 0$ as $i \rightarrow \infty$. Thus to show that (7.4.7) holds it only remains to show that $\|S_{h_i}^{k_i} \mathbf{x}_i - S(k_i h_i) \mathbf{x}_i\| \rightarrow 0$ as $i \rightarrow \infty$, but this follows from Proposition 3.6.8. Thus \mathcal{A}_0^+ is forward invariant.

A similar proof shows that \mathcal{A}_0^- is forward invariant under $S(\bullet)$. ■

We now suppose that the \mathcal{A}_h 's are backward invariant under the discretization for all $h \in (0, h_0)$ and show that this implies that \mathcal{A}_0^- and \mathcal{A}_0^+ are both backward invariant under the evolution of the underlying system (7.1.1). The proof of Proposition 7.4.6 is very similar to that of Proposition 7.4.5, except for the complications introduced by having to consider the backwards in time evolution of errors.

Proposition 7.4.6 *Suppose that (7.1.1) defines a dynamical system on \mathbb{R}^m with f locally Lipschitz and that this system is approximated numerically using a Runge-Kutta method (3.2.1–2), where for an implicit method the solution of (3.2.1–2) is as defined by Proposition 3.6.3. Then given a continuum of sets $\{\mathcal{A}_h\}_{h \in (0, h_0)}$ where the \mathcal{A}_h 's are uniformly bounded in \mathbb{R}^m and each \mathcal{A}_h is backward invariant under S_h the evolution operator for the discretized system, then with \mathcal{A}_0^- and \mathcal{A}_0^+ as defined in Definition 7.4.1 it follows that \mathcal{A}_0^+ and \mathcal{A}_0^- are both backward invariant under $S(\bullet)$, the evolution operator for the underlying system.*

Proof. First we show that \mathcal{A}_0^+ is backward invariant. As in the proof of Proposition 7.4.5, \mathcal{A}_0^+ is bounded and we can choose $\delta > 0$ such that $\mathcal{A}_h \subseteq \mathcal{N}(\mathcal{A}_0^+, \frac{1}{2}\delta)$ for all $h \in (0, h_0)$. Now since f is locally Lipschitz, $S(t)\mathbf{x}$ can only cease to exist if it becomes unbounded. Thus if \mathcal{A}_0^+ is not backward invariant we can choose $\mathbf{x}^* \in \mathcal{A}_0^+$ and $t^* > 0$ such that for all $\mathbf{x} \in \mathcal{N}(\mathcal{A}_0^+, \frac{1}{2}\delta)$ and all $-t \in [-t^*, 0]$ $S(-t)\mathbf{x}$ is well defined, $S(-t)\mathbf{x} \in \mathcal{N}(\mathcal{A}_0^+, \delta)$ and $S(-t^*)\mathbf{x}^* \notin \mathcal{A}_0^+$. With such a choice of \mathbf{x}^* and t^* we will derive a contradiction.

Since $\mathbf{x}^* \in \mathcal{A}_0^+$ there exists a monotonic decreasing sequence $\{h_i\}_{i=1}^\infty$ with $h_i \rightarrow 0$ as $i \rightarrow \infty$ and $\{\mathbf{x}_i\}_{i=1}^\infty$ with $\mathbf{x}_i \in \mathcal{A}_{h_i}$ and $\mathbf{x}_i \rightarrow \mathbf{x}^*$. Choose k_i to be the unique positive

integer such that $-k_i h_i \in [-t^*, -t^* + h_i)$. Now since \mathcal{A}_{h_i} is backward invariant there exists a negative orbit through \mathbf{x}_i , $\{S_{h_i}^{-n} \mathbf{x}_i\}_{n=1}^{\infty}$ with $S_{h_i}^{-n} \mathbf{x}_i \in \mathbf{x}_i$ for all $n \geq 0$. Note that this negative orbit need not be unique, but we will show below that any negative orbit has the properties that we require. We now show that

$$\|S_{h_i}^{-k_i} \mathbf{x}_i - S(-t^*) \mathbf{x}^*\| \rightarrow 0 \quad \text{as} \quad i \rightarrow \infty. \quad (7.4.8)$$

Since each \mathcal{A}_{h_i} is backward invariant $S_{h_i}^{-k_i} \mathbf{x}_i \in \mathcal{A}_{h_i}$ and then (7.4.8) implies that $S(-t^*) \mathbf{x}^* \in \mathcal{A}_0^+$, the contradiction that we require.

To show that (7.4.8) holds, note that

$$\begin{aligned} \|S_{h_i}^{-k_i} \mathbf{x}_i - S(-t^*) \mathbf{x}^*\| &\leq \|S_{h_i}^{-k_i} \mathbf{x}_i - S(-k_i h_i) \mathbf{x}_i\| + \|S(-k_i h_i) \mathbf{x}_i - S(-k_i h_i) \mathbf{x}^*\| \\ &\quad + \|S(-k_i h_i) \mathbf{x}^* - S(-t^*) \mathbf{x}^*\|. \end{aligned}$$

By construction $(t^* - k_i h_i) \rightarrow 0$ as $i \rightarrow \infty$ and hence $\|S(-k_i h_i) \mathbf{x}^* - S(-t^*) \mathbf{x}^*\| \rightarrow 0$ as $i \rightarrow \infty$. Since $\mathbf{x}_i \rightarrow \mathbf{x}$ by continuity with respect to initial data $\|S(-k_i h_i) \mathbf{x}_i - S(-k_i h_i) \mathbf{x}^*\| \rightarrow 0$ as $i \rightarrow \infty$. Thus to show that (7.4.8) holds it only remains to show that

$$\|S_{h_i}^{-k_i} \mathbf{x}_i - S(-k_i h_i) \mathbf{x}_i\| \rightarrow 0 \quad \text{as} \quad i \rightarrow \infty. \quad (7.4.9)$$

To show this we first show that for all $\mathbf{x} \in \mathcal{N}(\mathcal{A}_0^+, \frac{1}{2}\delta)$ and for any integers $n \geq 0$ and $h \geq 0$ such that $-nh \in [-t^*, 0]$

$$\|S_h^{-n} \mathbf{x} - S(-nh) \mathbf{x}\| \leq C_0 h^2 L \sum_{k=1}^n e^{kLh} \quad (7.4.10)$$

where C_0 is a positive constant and L is the Lipschitz constant for \mathbf{f} on $\mathcal{N}(\mathcal{A}_0^+, 2\delta)$.

We establish (7.4.10) by induction. First note that

$$\begin{aligned} \|S_h^{-1} \mathbf{x} - S(-h) \mathbf{x}\| &= \|S(-h)[S(h)S_h^{-1} \mathbf{x}] - S(-h)[S_h S_h^{-1} \mathbf{x}]\| \\ &\leq e^{hL} \|S(h)[S_h^{-1} \mathbf{x}] - S_h[S_h^{-1} \mathbf{x}]\| \end{aligned}$$

by Result 2.1.4. Hence by Proposition 3.6.7 $\|S_h^{-1} \mathbf{x} - S(-h) \mathbf{x}\| \leq e^{hL} C_0 h^2 L$ (where $C_0 = CM$ in (3.6.13)) and (7.4.10) holds for $n = 1$. To complete the induction suppose

(7.4.10) holds for $n = N$ and then

$$\begin{aligned}
\|S_h^{-(N+1)}\mathbf{x} - S(-(N+1)h)\mathbf{x}\| &\leq \|S(-h)[S(h)S_h^{-(N+1)}\mathbf{x}] - S(-h)[S_h^{-N}\mathbf{x}]\| \\
&\quad + \|S(-h)[S_h^{-N}\mathbf{x}] - S(-h)[S(-Nh)\mathbf{x}]\| \\
&\leq e^{hL}\|S(h)[S_h^{-(N+1)}\mathbf{x}] - S_h^{-1}[S_h^{-(N+1)}\mathbf{x}]\| \\
&\quad + e^{hL}\|S_h^{-N}\mathbf{x} - S(-Nh)\mathbf{x}\| \\
&\leq e^{hL}C_0h^2 + e^{hL}C_0h^2L\sum_{k=1}^Ne^{kLh} \\
&= C_0h^2L\sum_{k=1}^{N+1}e^{kLh}
\end{aligned}$$

where we have applied Result 2.1.4 and Proposition 3.6.7 again. This completes the inductive step and establishes (7.4.10).

Now (7.4.10) implies that

$$\begin{aligned}
\|S_{h_i}^{-k_i}\mathbf{x}_i - S(-k_ih_i)\mathbf{x}_i\| &\leq C_0h_i^2L\sum_{j=1}^{k_i}e^{jLh_i} \\
&\leq C_0h_i^2L\frac{e^{Lh_i}(e^{Lk_ih_i} - 1)}{e^{Lh_i} - 1} \\
&\leq C_0h_i^2L\frac{e^{Lh_i}(e^{Lk_ih_i} - 1)}{Lh_i} \\
&\leq C_0h_ie^{Lh_i}(e^{Lt^*} - 1)
\end{aligned}$$

since $k_ih_i \leq t^*$, and (7.4.9) follows. This completes the proof that \mathcal{A}_0^+ is backward invariant.

A similar proof shows that \mathcal{A}_0^- is backward invariant under $S(\bullet)$. ■

Remark A much neater proof of backward invariance which avoids the use of backwards in time error estimates has recently appeared in Hill and Süli [32].

Combining Propositions 7.4.5 and 7.4.6 we can show that if \mathcal{A}_h is invariant for each h and converges in the Hausdorff metric to a compact set \mathcal{A}_0 then \mathcal{A}_0 is an invariant set for the underlying system.

Theorem 7.4.7 *Suppose that (7.1.1) defines a dynamical system on \mathbb{R}^m and that this system is approximated numerically using a Runge-Kutta method (3.2.1–2), where for an implicit method the solution of (3.2.1–2) is as defined by Proposition 3.6.3. Then*

if $\{\mathcal{A}_h\}_{h \in (0, h_0)}$ is a continuum of sets where the \mathcal{A}_h 's are uniformly bounded in \mathbb{R}^m , each \mathcal{A}_h is invariant under S_h the evolution operator for the discretized system and \mathcal{A}_h converges to a compact set \mathcal{A}_0 in the Hausdorff metric as $h \rightarrow 0$ then \mathcal{A}_0 is invariant under $S(\bullet)$, the evolution operator for the underlying system.

Proof. By Theorem 7.4.4 $\mathcal{A}_0 = \mathcal{A}_0^- = \mathcal{A}_0^+$ and the result then follows from Propositions 7.4.5 and 7.4.6. ■

Remarks (i) This theorem is related to the results in Chapter 4. Theorem 4.2.8 shows that a continuous branch of fixed points either becomes unbounded or converges to a fixed point of the underlying system as $h \rightarrow 0$. We have now extended this result from fixed points to general invariant sets.

(ii) Even if \mathcal{A}_h does not converge in the Hausdorff metric as $h \rightarrow 0$ then \mathcal{A}_0^- and \mathcal{A}_0^+ are both invariant under $S(\bullet)$ and so the dynamics are still nonspurious. Note that we have not required that the invariant subsets of the system are isolated. For example if $f(y) \equiv 0$ then any subset of \mathbb{R}^m is invariant, which is one reason why we cannot guarantee convergence in the Hausdorff metric.

(iii) This theorem can be considered as a converse to theorems which prove that certain dynamics of the underlying system are preserved by the numerical approximation. For example, Beyn [3] and Eirola [14] prove that if (7.1.1) has a hyperbolic periodic orbit then a one-step discretization will also possess an invariant curve, for h sufficiently small, which converges to the periodic orbit of the underlying system as $h \rightarrow 0$. In contrast Theorem 7.4.7 implies that if the numerical approximation possesses an invariant curve for all h sufficiently small which converges to a closed loop as $h \rightarrow 0$ then this is an invariant curve for the underlying system.

For our final result we return to the case where the underlying system (7.1.1) possesses an attractor, but we no longer make the assumption that this attractor has the form $\mathcal{A} = \overline{W(\mathcal{E}^*)}$. Nevertheless we can show that if the numerical approximations \mathcal{A}_h converge in Hausdorff metric to a compact set \mathcal{A}_0 as $h \rightarrow 0$ then \mathcal{A}_0 is an invariant subset of \mathcal{A} which contains $\overline{W(\mathcal{E}^*)}$. Whilst if the \mathcal{A}_h do not converge in Hausdorff metric as $h \rightarrow 0$ then \mathcal{A}_0^- and \mathcal{A}_0^+ are both invariant subsets of \mathcal{A} which contain $\overline{W(\mathcal{E}^*)}$.

Theorem 7.4.8 Suppose that (7.1.1) defines a dynamical system on \mathbb{R}^m with an attractor \mathcal{A} which attracts a bounded forward invariant neighbourhood N of itself. Let \mathcal{E}^*

be the set of hyperbolic fixed points contained in \mathcal{A} , and $W(\mathcal{E}^*)$ the union of the unstable manifolds of these hyperbolic fixed points. Suppose that f is C^1 on a neighbourhood of \mathcal{E}^* . If this system is approximated numerically using a Runge-Kutta method (3.2.1–2), where for an implicit method the solution of (3.2.1–2) is as defined by Proposition 3.6.3, then there exists $h_0 > 0$ such that for $h \in (0, h_0)$ the numerical solution defines a continuous discrete dynamical system on $N_h \supseteq N$, which possesses an attractor \mathcal{A}_h that attracts N_h (and hence N), and, with \mathcal{A}_0^- and \mathcal{A}_0^+ as defined in Definition 7.4.1, satisfies

$$\overline{W(\mathcal{E}^*)} \subseteq \mathcal{A}_0^- \subseteq \mathcal{A}_0^+ \subseteq \mathcal{A}.$$

Here \mathcal{A}_0^- and \mathcal{A}_0^+ are both invariant under $S(\bullet)$ the evolution operator of the underlying system, and if $\{\mathcal{A}_h\}_{h \in (0, h_0)}$ converges in Hausdorff metric to a compact set \mathcal{A}_0 as $h \rightarrow 0$ then $\mathcal{A}_0 = \mathcal{A}_0^- = \mathcal{A}_0^+$ and hence $\overline{W(\mathcal{E}^*)} \subseteq \mathcal{A}_0 \subseteq \mathcal{A}$ and \mathcal{A}_0 is also invariant under $S(\bullet)$.

Proof. Proposition 7.2.1 implies that the numerical solution defines a continuous discrete dynamical system with an attractor \mathcal{A}_h for $h \in (0, h_0)$ for some $h_0 > 0$. Then Proposition 7.2.1 together with Proposition 7.4.2 imply that $\mathcal{A}_0^+ \subseteq \mathcal{A}$.

On noting that $\overline{W_h(\mathcal{E}^*)} \subseteq \mathcal{A}_h$ so that $\text{dist}(\overline{W(\mathcal{E}^*)}, \mathcal{A}_h) \leq \text{dist}(\overline{W(\mathcal{E}^*)}, \overline{W_h(\mathcal{E}^*)})$, Proposition 7.3.2 together with Proposition 7.4.3 implies that $\overline{W(\mathcal{E}^*)} \subseteq \mathcal{A}_0^-$.

The invariance of \mathcal{A}_0^- and \mathcal{A}_0^+ under $S(\bullet)$ follows from Proposition 7.4.6. Finally in the case where $\{\mathcal{A}_h\}_{h \in (0, h_0)}$ converge in Hausdorff metric to a compact set \mathcal{A}_0 the equality of \mathcal{A}_0 , \mathcal{A}_0^- and \mathcal{A}_0^+ follows from Theorem 7.4.4. ■

Remarks (i) In line with Remark(vi) after Theorem 7.3.3 we can easily extend this result to show that $\overline{W(\mathcal{E}^*) \cup \mathcal{P}^*} \subseteq \mathcal{A}_0^-$ and it may be possible to further extend the result to show that $\overline{W(\mathcal{E}^*) \cup W(\mathcal{P}^*)} \subseteq \mathcal{A}_0^-$. (ii) In the case where $\{\mathcal{A}_h\}_{h \in (0, h_0)}$ does not converge in Hausdorff metric as $h \rightarrow 0$ we have considered the properties of $\mathcal{A}_0^- = \liminf_{h \rightarrow 0} \mathcal{A}_h$ and $\mathcal{A}_0^+ = \limsup_{h \rightarrow 0} \mathcal{A}_h$. An alternative approach would be to consider convergent subsequences of \mathcal{A}_{h_i} . This approach is followed by Hill and Süli [32], who, in the context of the Hale, Lin and Raugel results, prove that every sequence $\{\mathcal{A}_{h_i}\}_{i=1}^\infty$ with $h_i \rightarrow 0$ contains a convergent subsequence converging in Hausdorff metric to an invariant subset of \mathcal{A} .

Finally note that throughout this chapter we have considered the convergence of numerical invariant sets to invariant sets of the underlying system. To do this we

have simply thought of invariant sets and attractors as subsets of \mathbb{R}^m , and have only considered the dynamics of the underlying system and its numerical approximations in order to establish properties of and relationships between these sets. We have run out of space and time before we have even begun to address the important task of comparing the dynamics of the underlying system on an invariant set with the dynamics of its numerical approximation on an approximating numerical invariant set.

Bibliography

- [1] J-P. Aubin and A. Cellina. *Differential Inclusions*. Springer-Verlag, 1984.
- [2] J-P. Aubin and H. Frankowska. *Set-Valued Analysis*. Birkhäuser, 1990.
- [3] W.J. Beyn. On closed invariant curves for one step methods. *Numer. Math.*, 51:103–122, 1987.
- [4] W.J. Beyn. On the numerical approximation of phase portraits near stationary points. *SIAM J. Num. Anal.*, 24:1095–1113, 1987.
- [5] W.J. Beyn. Numerical methods for dynamical systems. In W.A. Light, editor, *Advances in Numerical Analysis; Volume 1*. Clarendon Press, Oxford, 1991.
- [6] K. Burrage and J.C. Butcher. Stability criteria for implicit Runge-Kutta processes. *SIAM J. Num. Anal.*, 16:46–57, 1979.
- [7] J.C. Butcher. Implicit Runge-Kutta processes. *Math. Comp.*, 18:50–64, 1964.
- [8] J.C. Butcher. A stability property of implicit Runge-Kutta methods. *BIT*, 15:358–361, 1975.
- [9] J.C. Butcher. *The Numerical Analysis of Ordinary Differential Equations: Runge-Kutta and General Linear Methods*. Wiley, Chichester, 1987.
- [10] G. Dahlquist. A special stability problem for linear multistep methods. *BIT*, 3:27–43, 1963.
- [11] G. Dahlquist. Error analysis for a class of methods for stiff non-linear initial value problems. In *Numerical Analysis, Dundee 1975*, pages 60–74. Springer, 1975.
- [12] G. Dahlquist. G-stability is equivalent to A-stability. *BIT*, 18:384–401, 1978.

- [13] K. Dekker and J.G. Verwer. *Stability of Runge-Kutta Methods for Stiff Nonlinear Equations*. North Holland, Amsterdam, 1984.
- [14] T. Eirola. Invariant curves of one step methods. *BIT*, 28:113–122, 1988.
- [15] T. Eirola and O. Nevanlinna. What do multistep methods approximate ? *Numer. Math.*, 53:559–569, 1988.
- [16] C.M. Elliott. The Cahn-Hilliard model for the kinetics of phase separation. In J.F. Rodrigues, editor, *Mathematical models for phase change problems*. Birkhäuser-Verlag, 1989.
- [17] C.M. Elliott and A.M. Stuart. The global dynamics of discrete semilinear parabolic equations, 1992. To appear in *SIAM J. Num. Anal.*
- [18] C. Foias, M.S. Jolly, I.G. Kevrekidis, and E.S. Titi. Dissipativity of numerical schemes. *Nonlinearity*, 4:591–613, 1991.
- [19] C. Foias and E.S. Titi. Determining nodes, finite difference schemes and inertial manifolds. *Nonlinearity*, 4:135–153, 1991.
- [20] D.A. French and S. Jensen. Long time behaviour of arbitrary order continuous time galerkin schemes for some one-dimensional phase transition problems, 1992. Preprint.
- [21] D.H. Griffel. *Applied Functional Analysis*. Ellis Horwood, Chichester, 1981.
- [22] D.F. Griffiths. The dynamics of some linear multistep methods with step-size control. In *Proceedings of the 12th Biennial Dundee Conference on Numerical Analysis*. Pitman, London, 1987.
- [23] D.F. Griffiths, P.K. Sweby, and H.C. Yee. Spurious steady state solutions of explicit Runge-Kutta schemes. *IMA J. Num. Anal.*, 12:319–338, 1992.
- [24] J. Guckenheimer and P. Holmes. *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*. Applied Mathematical Sciences 42. Springer-Verlag, New York, 1983.
- [25] E. Hairer, A. Iserles, and J.M. Sanz-Serna. Equilibria of Runge-Kutta methods. *Numer. Math.*, 58:243–254, 1990.

- [26] E. Hairer, S.P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Springer-Verlag, Berlin, 1987.
- [27] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Springer-Verlag, Berlin, 1991.
- [28] J.K. Hale. *Asymptotic Behaviour of Dissipative Systems*. American Mathematical Society, 1980.
- [29] J.K. Hale, X.-B. Lin, and G. Raugel. Upper semicontinuity of attractors for approximations of semigroups and partial differential equations. *Math. Comp.*, 50:89–123, 1988.
- [30] J.K. Hale and G. Raugel. Lower semicontinuity of attractors of gradient systems and applications. *Ann. Mat. Pura ed Applicata*, IV:281–326, 1989.
- [31] P. Hartman. *Ordinary Differential Equations*. Wiley, 1964.
- [32] A.T. Hill and E. Süli. Set convergence for discretizations of the attractor. Report 93/8, Oxford University Computing Laboratory, 1993.
- [33] A.T. Hill and E. Süli. Upper semicontinuity of attractors for linear multistep methods approximating sectorial evolution equations. Report 92/8, Oxford University Computing Laboratory, 1993.
- [34] M.W. Hirsch and S. Smale. *Differential Equations, Dynamical Systems and Linear Algebra*. Academic Press, London, 1974.
- [35] A.R. Humphries. Nonlinear stability theory for ODEs, an asymptotic approach, 1993. In Preparation.
- [36] A.R. Humphries and A.M. Stuart. Runge-Kutta methods for dissipative and gradient dynamical systems, 1992. To appear in *SIAM J. Num. Anal.*
- [37] W.H. Hundsdorfer and M.N. Spijker. A note on B-stability of Runge-Kutta methods. *SIAM J. Num. Anal.*, 24:583–594, 1981.
- [38] A. Iserles. Stability and dynamics of numerical methods for nonlinear ordinary differential equations. *IMA J. Num. Anal.*, 10:1–30, 1990.

- [39] A. Iserles, A.T. Peplow, and A.M. Stuart. A unified approach to spurious solutions introduced by time discretization. Part I: Basic theory. *SIAM J. Num. Anal.*, 28:1723–1751, 1991.
- [40] A. Iserles and A.M. Stuart. A unified approach to spurious solutions introduced by time discretization. Part II: BDF-like methods. *IMA J. Num. Anal.*, 12:487–502, 1992.
- [41] Urs Kirchgraber. Multi-step methods are essentially one-step methods. *Numer. Math.*, 48:85–90, 1986.
- [42] P.E. Kloeden and J. Lorenz. Stable attracting sets in dynamical systems and their one-step discretizations. *SIAM J. Num. Anal.*, 23:986–995, 1986.
- [43] G.J. Lord. The dynamics of numerical methods for initial value problems, 1993. Phd thesis, in preparation.
- [44] E.N. Lorenz. Deterministic nonperiodic flow. *J. Atmospheric Sci.*, 20:130–141, 1963.
- [45] M.Z. Lui and J.F.B.M. Kraaijevanger. On the solvability of equations arising in implicit Runge-Kutta methods. *BIT*, 28:825–838, 1988.
- [46] A.C. Newell. Finite amplitude instabilities of partial differential equations. *SIAM J. Appl. Math.*, 33:133–160, 1977.
- [47] J.M. Sanz-Serna. Runge-Kutta schemes for Hamiltonian systems. *BIT*, 28:877–883, 1988.
- [48] J.M. Sanz-Serna. Numerical ordinary differential equations versus dynamical systems. In D.S. Broomhead and A. Iserles, editors, *The dynamics of numerics and the numerics of dynamics*, Oxford, 1992. Clarendon Press.
- [49] J.M. Sanz-Serna and A.M. Stuart. A note on uniform in time error estimates for approximations to reaction diffusion equations. *IMA J. Num. Anal.*, 12:457–462, 1992.
- [50] H.J. Stetter. *Analysis of Discretisation Methods for Ordinary Differential Equations*. Springer, New York, 1973.

- [51] I. Stewart. *Does God Play Dice ? The New Mathematics of Chaos*. Penguin, London, 1990.
- [52] D. Stoffer and K. Nipp. Invariant curves for variable step size integrators. *BIT*, 31:169–180, 1991.
- [53] A.M. Stuart. The global attractor under discretization. In D. Roose, B. De Dier, and A. Spence, editors, *Continuation and Bifurcations: Numerical Techniques and Applications*. Kluwer Academic Publishers, 1990.
- [54] A.M. Stuart. Numerical analysis of dynamical systems. In *Acta Numerica*. Cambridge University Press, 1994.
- [55] A.M. Stuart and A.R. Humphries. The essential stability of local error control for dynamical systems, 1992. To appear in *SIAM J. Num. Anal.*
- [56] A.M. Stuart and A.R. Humphries. Model problems in numerical stability theory for initial value problems, 1992. To appear in *SIAM Review*.
- [57] A.M. Stuart and A.T. Peplow. The dynamics of the theta method. *SIAM J. Sci. Stat. Comp.*, 12:1351–1372, 1991.
- [58] R. Temam. *Infinite Dimensional Dynamical Systems in Mechanics and Physics*. Springer, New York, 1989.
- [59] S. Wiggins. *Introduction to Applied Nonlinear Dynamical Systems and Chaos*. Springer-Verlag, New York, 1990.